Semantic search using augmented retrieval generation: application of Artificial Intelligence in Brazilian broadcasting regulatory statutes

João Alberto de Oliveira Lima
Universidade de Brasília
Senado Federal
joaolima@ccom.unb.br

Lauro César Araujo
Universidade de Brasília
Senado Federal
laurocesar@ccom.unb.br

BIOGRAPHIES

João Alberto de Oliveira Lima is a legislative IT analyst at the Federal Senate, where he has worked since 1995 leading the Lexml.gov.br and Normas.leg.br projects. He holds a PhD in Information Science and a PhD in Law, both from the University of Brasília. Researcher affiliated to Research Center on Communication Policy, Law, Economics, and Technology (CCOM/UnB).

Lauro César Araujo is a specialist in Information Architecture and Ontology, has a PhD in Information Science from University of Brasília and is also a legislative IT analyst at the Federal Senate of Brazil. Researcher affiliated to Research Center on Communication Policy, Law, Economics, and Technology (CCOM/UnB).

STRUCTURED ABSTRACT

Purpose: The paper aims to introduce a novel approach for searching legislative texts using multi-layered embeddings, aiming enhance the accuracy and efficiency of legal information retrieval using legal knowledge representation at various granularities.

Methodology/approach/design: The approach involves experimentation in creating embeddings for individual articles, their components, and structural groupings. It uses Retrieval Augmented Generation (RAG) models to provide accurate responses tailored to the user's query, exploring concepts of aboutness, semantic chunking, and hierarchy within legal texts.

Findings: The method enhances both the accuracy and efficiency of legal information retrieval. It demonstrates how RAG can improve semantic research over structured legal documents, focusing on Brazilian Constitution and Brazilian broadcasting regulatory statutes.

Practical implications: The research suggests changes to incorporate multi-layered embeddings and RAG models, impacting legal practice, policy making, and regulatory compliance. Future information systems can be built with a focus on findability by design, which can profoundly change the user experience and clarity of understanding of regulatory statutes.

Originality/value: The novelty of this paper lies in its introduction of a multi-layered embeddings approach combined with RAG models for searching legislative texts. This method offers a more granular and potentially more effective way to retrieve legal information compared to traditional methods that focus solely on entire documents. The focus on Brazilian broadcasting regulatory statutes provides a valuable example of how this approach can be applied in a real-world setting. This research can be beneficial to legal professionals, policy makers, AI researchers, and the public by facilitating faster and more precise navigation of legal documents.

Keywords (Required, 5 KEYWORDS separated by periods as follows)

Multi-layered embeddings-based retrievel. Legislative Texts. Retrieval Augmented Generation (RAG). Semantic Search. Brazilian Broadcasting Regulatory Statutes.

INTRODUCTION

The increasing volume and complexity of legal information pose significant challenges for legal professionals and researchers seeking relevant knowledge. Traditional keyword-based search often falls short in capturing the nuances of legal language and the intricate relationships within legal documents (Saravanan et al., 2009; Mimouni, et al. 2014).

Recent advancements in generative AI and retrieval augmented generation (RAG) offer promising avenues for more efficient and accurate legal information retrieval. Embeddings — compact vector text representations — effectively the meanings of words, phrases, or documents (Yu & Dredze, 2015; Chersoni et al, 2021). Current methods often embed whole documents, articles, or text fragments. Legislative documents, including bills and legal normative statutes, inherently have a hierarchical structure, with articles composed of paragraphs, clauses and further subdivisions, often grouped into chapters and sections under broader titles. This intrinsic hierarchy suggests a need for a more granular approach to representing legal knowledge by means of embeddings.

This paper proposes a multi-layered embedding-based retrieval method that captures the semantic content of legal texts at various levels of granularity. By creating embeddings for articles, their components, and their structural groupings, we aim to provide a more nuanced and comprehensive representation of legal knowledge. Our proposed approach enables RAG models to respond to user queries with varying levels of detail, ranging from small portions to comprehensive sections.

FUNDAMENTAL CONCEPTS

Before we proceed with the exploration of our proposed method and practical examples, it is essential to lay a solid foundation by clarifying the core concepts that are fundamental to this research. These core concepts are important because they lay the groundwork for our methodology and ensure that our discussion is understandable not only to IT analysts but also to various stakeholders in the legislative field, who may not be intimately familiar with the terminologies discussed herein. By elucidating these key concepts, our goal is to eliminate any potential knowledge gaps, thereby fostering a deeper understanding of the advanced techniques utilized to enhance the retrieval of legal information.

Embeddings: At the heart of our approach lie embeddings, dense vector representations of text that encapsulate semantic meanings and relationships within documents. These vectors position words, sentences, or entire documents[1] in a multidimensional space, where the spatial proximity between points signifies semantic similarity. For instance, within the legal sphere, embeddings can discern the nuanced differences in meaning between terms like “contract” and “agreement” based on contextual usage, offering a refined perspective on legal terminology.

Furthermore, it is essential to emphasize the importance of analyzing the content undergoing embedding processes. Beyond capturing the contextual meaning of text, embeddings also reflect some syntactic features of documents. Texts that are semantically identical but presented with slight formal variations—such as differences in capitalization or line breaks—result in distinct vectors. This attention to semantic and syntactic features shows how embeddings comprehensively capture a wide range of linguistic nuances.

Moreover, it’s important to challenge the conventional reliance on keyword-based retrieval methods such as TF/IDF (Ayed & Biskri, 2020), which primarily focus on the frequency of term occurrences. Unlike these traditional techniques, embeddings provide a more nuanced approach by concentrating on the semantic content of texts. This allows them to discern the underlying meanings and themes, rather than merely counting words. This semantic focus makes embeddings particularly effective in complex fields like law, where the exact terms used may vary but the overarching legal concepts and relationships remain consistent.

Before exploring further into the other concepts, one might wonder about the intrinsic nature of an embedding. How does it capture and reflect the essence of the text it represents? This question leads us to a pivotal realization: embeddings can be seen as the digital “aboutness” of the content they encode. By distilling complex textual information into a dense vector form, embeddings provide a quantifiable representation of the subject matter, themes, and relationships inherent in the text. This understanding paves the way for leveraging embeddings as a tool for navigating the semantic landscape of legislative texts.

Aboutness: The concept of aboutness refers to the main theme or subject matter addressed by a text. It encompasses the entities, concepts, and relationships discussed within, providing a comprehensive overview of its content. Identifying a document’s aboutness is necessary for accurately pinpointing relevant information in response to specific queries, thereby facilitating the precise retrieval of legal information.

[1] When dealing with texts that exceed the maximum token limit of the language model used for embedding, it becomes necessary to segment the text into semantic chunks. Each chunk is then processed to calculate its embedding. To represent the entire text, the arithmetic mean of these embedding vectors is computed, ensuring a cohesive representation despite the token limitation.

In examining the foundational concept of 'aboutness' that underpins our multi-layered embedding approach for legal texts, it is valuable to reference Gilbert Ryle's early analysis of the term in his article "About," from Analysis in 1933. Ryle differentiates among interpretations of 'aboutness', particularly emphasizing the 'about conversational' aspect. He articulates that an expression captures the central theme of a conversation if a considerable number of sentences within that conversation either directly use that expression, employ a synonym, or make an indirect reference to the concept it denotes. This occurrence should dominate without a competing expression or referent being more prevalent across the dialogue. Ryle's viewpoint suggests that 'aboutness' serves as a marker for the primary concern within a discourse, indicated through frequent references or mentions, rather than through the mere presence of specific keywords (Ryle, 1933, p. 11).

This consideration of Ryle's perspective aids in refining our approach to embedding legal documents. Recognizing 'aboutness' as more than a function of explicit text, but rather as a reflection of the thematic consistency across a document, informs our strategy. It acknowledges that the core theme or subject matter of a discourse is not necessarily tied to the explicit repetition of certain terms but is instead often conveyed through varied linguistic mechanisms that signal ongoing engagement with the theme.

Chunking: This process involves dividing text into smaller, manageable segments, such as sentences, paragraphs, or specific legal clauses. Segmenting complex documents into these constituent parts allows for a granular examination of each segment's content and its contribution to the document's overarching meaning, facilitating the analysis and processing of intricate legal texts.

Semantic chunking, as currently practiced in semchunk[2], offers a method for dividing text into smaller, semantically coherent segments, utilizing a recursive division technique that initially splits text using contextually meaningful separators. This process repeats until all segments reach a specific size, ensuring each chunk represents a unique thematic or conceptual element. However, this approach overlooks the intrinsic hierarchical organization of legislative texts. These texts are not merely collections of semantically coherent segments; they are structured in a complex hierarchy. While traditional semantic chunking effectively creates manageable segments, it overlooks the hierarchical relationships. It treats all segments as if they are on the same level, disregarding the structured organization that defines their legal significance and their interrelations.

Retrieval Augmented Generation (RAG) is an advanced method that enhances response formulation by leveraging data from databases as a foundational source. This technique capitalizes on both traditional access methods and the semantic prowess of embeddings, thus ensuring that the responses generated are not only relevant but also contextually rich. By integrating RAG into the system, it becomes possible to craft answers that are deeply informed by the underlying data, offering a more nuanced understanding to users' queries.

A clear grasp of the aforementioned concepts is essential before exploring the specific characteristics of articulated normative texts. The ensuing section will elaborate on how these principles are applied to represent the complex, hierarchical nature of legislative texts, enabling more effective and nuanced retrieval of legal knowledge.

THE SPECIFICITY OF LEGAL INFORMATION

Legislative texts possess a unique characteristic: an inherent hierarchical structure. This structure reflects the systematization of legal dispositions, ranging from broad legal areas to specific articles, clauses, and sub-clauses. This inherent hierarchy provides a natural framework for representing legal knowledge with varying levels of granularity.

In structured normative texts such as constitutions, every detail has substantial semantic importance. Unlike ordinary language, where redundancy and informality are common, constitutions are crafted documents where each word, phrase, and clause is deliberately chosen to convey specific legal meaning and effect. This precision necessitates a granular approach to information retrieval. Analyzing and representing legal knowledge at varying levels of detail – from individual clauses and sub-clauses to broader articles, chapters and titles – allows us to capture the full spectrum of meaning embedded within these texts.

Figure 1 depicts the organizational structure of legislative documents in Brazil, conforming to the guidelines of Complementary Law No. 95/1998. At the highest level is the "Legal Standard" document (*Norma Jurídica*), under which the Main Text (*Texto Principal*) is the primary body of the document. Optionally, an Annex Text (*Texto Anexo*) may be included, indicated by a dashed line, suggesting its non-mandatory nature. It's important to note that the number of Annex Texts can vary, ranging from none to one or several, indicating the flexibility in the inclusion of such texts.

The foundational unit of this framework is the "Article," consisting of a mandatory "Lead paragraph" ("*caput*") and, potentially, one or more "Paragraphs" (*Parágrafo*), indicated by dashed lines to denote their optional status.[3] These "Paragraphs" elucidate

[2] https://github.com/umarbutler/semchunk, accessed in 2024-06-12.

[3] In the context of U.S. legislation, the term 'Articles' typically refers to the seven divisions of the U.S. Constitution that lay out the frame of the federal government. Unlike in Brazilian law, where 'Articles' serve as fundamental units within legislative

details or exceptions pertaining to the "Lead paragraph." If enumeration is necessary, the lead paragraph or paragraphs can branch into "Sections" (*incisos*), which may be further divided into "Items" (*alíneas*), and then into "Subitems" (*item*) for detailed specificity.

Simple statutes may consist only of articles, but complex statutes and codes group articles into progressively larger units: "Sections [group of Articles]" and "Subsections" are possible components of "Chapters," which are compiled under "Titles." These titles are then organized into "Books," and multiple books can be part of a larger "Part," reflecting the elaborate nature of legal codes.

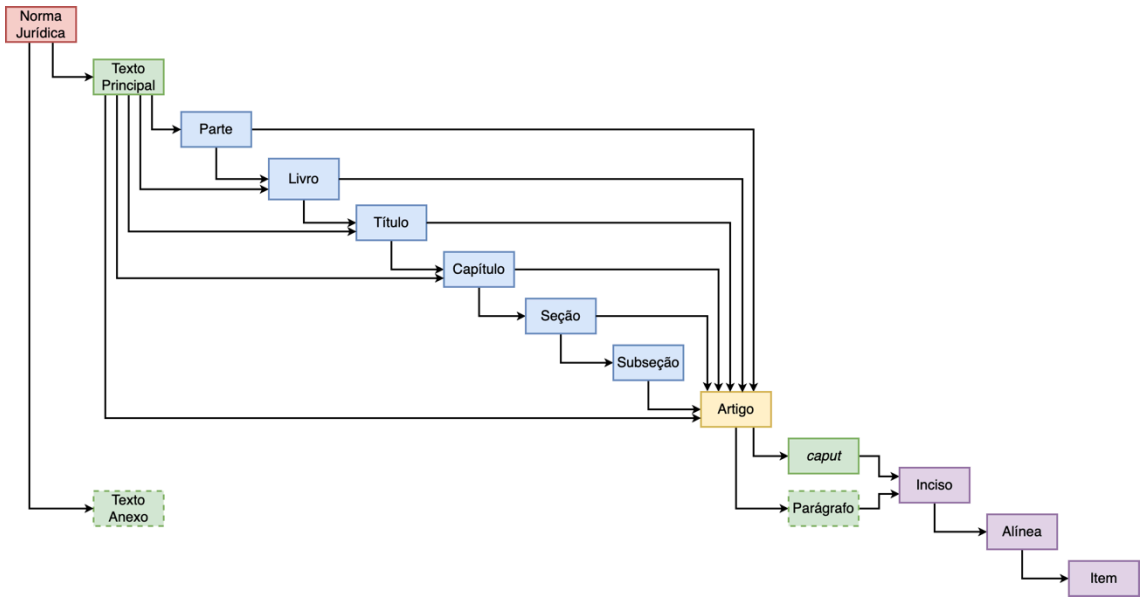


Figure 1. Hierarchy of legislative documents in Brazil.

The Figure 2 provides an excerpted view of the hierarchical structure of Title I and Chapter I of Title II of the Brazilian Constitution. It visually represents how the constitutional text is organized, starting from the broadest divisions and narrowing down to specific provisions. Each "Article" serves as a foundational element, with the "Lead paragraph" (*caput*) stating the principal point. "Sole Paragraphs" provide further explanations or stipulations when an article has only one. For detailed enumeration within an article, "Sections" (*incisos*) break down the content into subsections, each identified by Roman numerals. These sections can further subdivide into "Items" (*alíneas*), denoted by lowercase letters, and 'Subitems' for the most specific details, though not depicted in this portion of the structure. This systematic approach allows for a clear and organized reading of the constitutional provisions.

texts, in U.S. law, 'Sections' within codes and statutes are the comparable unit of division that detail specific provisions of the law. Therefore, what is termed an 'Article' (*Artigo*) in Brazil could often be likened to a 'Section' in U.S. legislation.

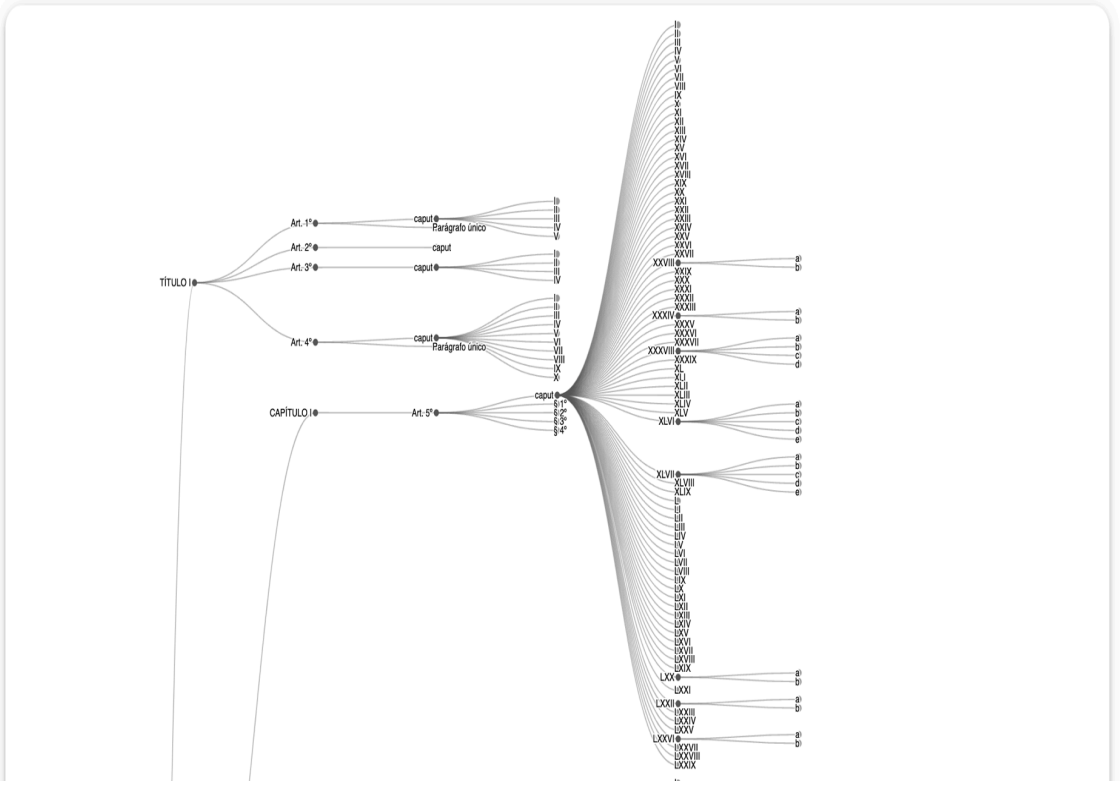


Figure 2. Segment of the Hierarchical Structure of the Brazilian Constitution

The Figure 3 is an extract from Title I of the Brazilian Constitution, highlighting its hierarchical structure through color-coded annotations. The yellow color highlights the beginning of the main text of the Brazilian Constitution, while the red border frames the entirety of Title I. The blue highlight focuses on the "Lead paragraph" (*caput*) of Article 1, which declares Brazil as an indivisible federation and defines its democratic, legal foundation. The orange highlight designates the entirety of Article 1, detailing the core values of the federation such as sovereignty, citizenship, human dignity, social values of labor and free enterprise, and political pluralism. Lastly, the green highlight specifically illuminates Section III (*inciso*) of the lead paragraph of Article 1, underscoring 'human dignity' as a fundamental principle.

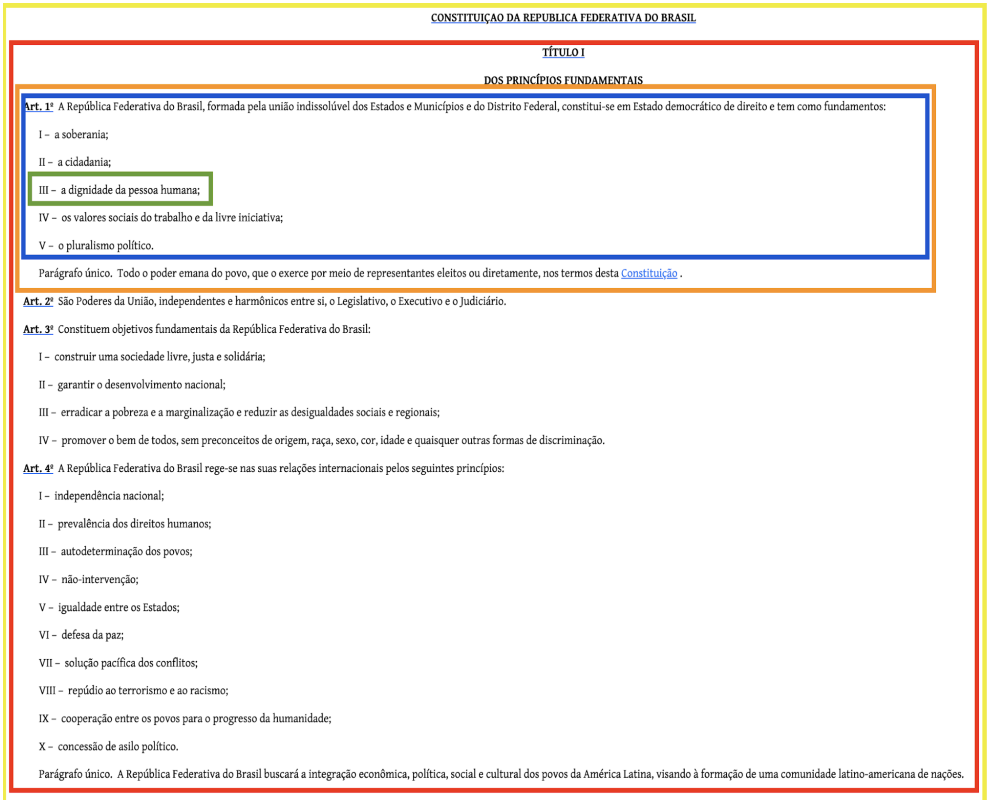


Figure 3. Structural Elements of Title I of the Brazilian Constitution

Considering their fundamental role in the systematization of legal norms and principles, "Articles" emerge as the most suitable unit for initial semantic chunking within legislative texts. As illustrated in Figure 3, applying this approach to Title I of the Brazilian Constitution would result in four distinct chunks, each representing an individual article. While this provides a solid basis for analysis, our proposed methodology, detailed in subsequent sections, seeks to further enhance this process. Generating a broader set of chunks for embedding allows us to create a more detailed and complete representation of legal knowledge, ultimately leading to greater accuracy and effectiveness in legal information retrieval.

PROPOSAL FOR UTILIZING EMBEDDINGS WITH VARIABLE GRANULARITY: A MULTI-LAYERED APPROACH

This section further explores the proposed multi-layered embedding-based retrieval approach, outlining its various levels and investigating potential avenues for future development.

- Document-Level Embeddings: At the highest level, we generate embeddings for the entire legislative document. This captures the overarching theme, purpose, and scope of the legal text.
- Text Component-Level of Documents Embeddings: This level concentrates on capturing the essence of document components' texts, which can be systematically presented through articles (such as the main text) or depicted using various structures (like tables or unstructured text). Examples encompass the main text and additional texts such as justifications (in bills) and schedules (in annexes), with each being assigned its own embedding to represent its distinct contribution to the overarching legal context.
- Article-Level Embeddings: Each article, as a fundamental unit of legal text, receives its own embedding. This captures the specific legal issue addressed by the article and its core provisions.
- Article Component-Level Embeddings: Further granularity is achieved by creating embeddings for individual paragraphs, clauses, and even sub-clauses within each article. This allows for a nuanced understanding of the specific details and nuances of each legal provision.
- Grouping-Level Embeddings: Embeddings are also generated for titles, chapters, and sections. This captures the broader themes and relationships between groups of articles.

By incorporating embeddings at these different levels, we create a multi-layered representation of legal knowledge, enabling RAG models to respond to user queries with varying levels of detail.

To illustrate the granularity achieved by our proposed method, let's consider the example presented in Figure 3. A standard approach using articles as the embedding unit would segment this text into 4 chunks, corresponding to the 4 articles present. However, our hierarchical approach yields a much richer representation. At the Document-Level, we have 1 chunk encompassing the entire text. Similarly, the Text Component-Level also results in 1 chunk, as it encompasses the main body of the text in this example. Moving to the Article-Level, we retain the 4 chunks corresponding to the articles. However, the Article Component-Level further breaks these down into 25 chunks: 4 lead paragraphs (*caputs*), 2 sole paragraphs, and 19 sub-sections (*incisos*). Finally, the Grouping-Level provides 1 chunk representing the single title. In total, our method generates 32 chunks, offering a significantly more detailed and nuanced representation compared to the 4 chunks of the traditional approach.

The Figure 3 not only showcases the structure of Title I of the Brazilian Constitution but also reveals an insightful detail regarding Article Component-Level embedding within legal texts. Specifically, when it comes to enumerative elements such as 'Sections' (*incisos*), 'Items' (*alíneas*), and 'Subitems' (*itens*), it's not sufficient to consider their text in isolation. Since each element elaborates on a part of a larger context, it's essential to embed their text within the context of the superior elements up to the level of the 'Lead paragraph' (*caput*) of the article or the paragraph to which they belong. For instance, instead of isolating the phrase "*a dignidade da pessoa humana*" (the dignity of the human person), the embedded text should read as follows: "[*A República Federativa do Brasil, formada pela união indissolúvel dos Estados e Municípios e do Distrito Federal, constitui-se em Estado democrático de direito e tem como fundamentos:*] *dignidade da pessoa humana*"[4]. This method of embedding ensures that the specific provisions are always understood within the full scope of their intended legal framework.

COMPARATIVE VISUALIZATION OF TRADITIONAL VS MULTI-LAYERED EMBEDDING-BASED RETRIEVAL APPROACHES

This section provides a graphical comparison between traditional segmentation methods and the proposed a multi-layered embedding-based retrieval approach in the context of the current Brazilian Constitution text. Through an innovative experiment, we employed the text-embedding-3-large model, configured to generate 256-dimensional vectors. To present the data in a comprehensible manner, we reduced these high-dimensional embeddings to a two-dimensional plane using the dimensionality reduction technique known as pacmap[5]. The visualizations were then created using the interactive graphing library Plotly[6], facilitating a clearer understanding of the data structures and their semantic correlations.

The Figure 4 exemplifies our multi-layered embedding approach juxtaposed with the traditional method, using the same corpus of the Brazilian Constitution text for both. The visualization brings to light the enhanced ability of the hierarchical method to capture and represent the multifaceted semantic nuances of legal texts, as compared to the more uniform view provided by the traditional method. Figure 4 showcases a visualization of the embeddings of the Brazilian Federal Constitution's provisions, displayed in a two-dimensional space for easier comprehension. This visualization is divided into four distinct quadrants, each providing insights into the hierarchical structure and the semantic relationships of the text, based on the embeddings.

The first quadrant focuses exclusively on the embeddings of the articles. Here, each point represents an article, plotted according to its embedding in a simplified two-dimensional space. This offers a broad overview of how each article is positioned relative to others, potentially reflecting their semantic proximity and thematic linkages. Moving to the second quadrant, the visualization becomes significantly more intricate. It includes embeddings from various levels of granularity, as advocated in the presented article. This comprehensive approach incorporates not only the articles but also their components such as paragraphs, clauses, and sub-clauses, as well as their groupings like titles and chapters. The result is a rich tapestry of points, each signifying a specific segment of the constitutional text, hence revealing the complex interrelations and the semantic depth captured by the multi-layered embedding-based retrieval process.

The third quadrant provides a zoomed-in view focusing on the vicinity of Article 5, spotlighting the singular embedding of this particular article. In contrast, the fourth quadrant extends this closer look by including all embeddings associated with Article 5,

[4] "[The Federative Republic of Brazil, formed by the indissoluble union of the states and municipalities and of the Federal District, is a legal democratic state and is founded on:] the dignity of the human person;"

[5] Pacmap (https://github.com/YingfanWang/PaCMAP) is a dimensionality reduction algorithm that preserves the global data topology, ensuring that the relative distances in the high-dimensional space are meaningfully maintained in the reduced space, thus allowing for a more accurate representation of complex data structures.

[6] Plotly (https://plotly.com/python/) is an interactive, open-source graphing library for Python that enables users to create visually appealing, dynamic, and interactive visualizations. Plotly supports a range of charts and diagrams, including line charts, scatter plots, and complex three-dimensional models, which are essential for data analysis and presentation.

such as its numerous clauses and sub-clauses. Given that Article 5 contains over 70 such components, this quadrant vividly demonstrates the increased density of points surrounding the Article 5 embedding. It illustrates the detailed semantic landscape that emerges when each subdivision is given a distinct representation, underscoring the sophistication of the proposed multi-layered embedding approach.

This method's capability to represent legal texts at different levels of granularity ensures that whether a query seeks an overview or a detailed examination, the retrieval system can provide to the LLM the relevant chunks as database source to create an answer that matches the user's specific request with remarkable precision.

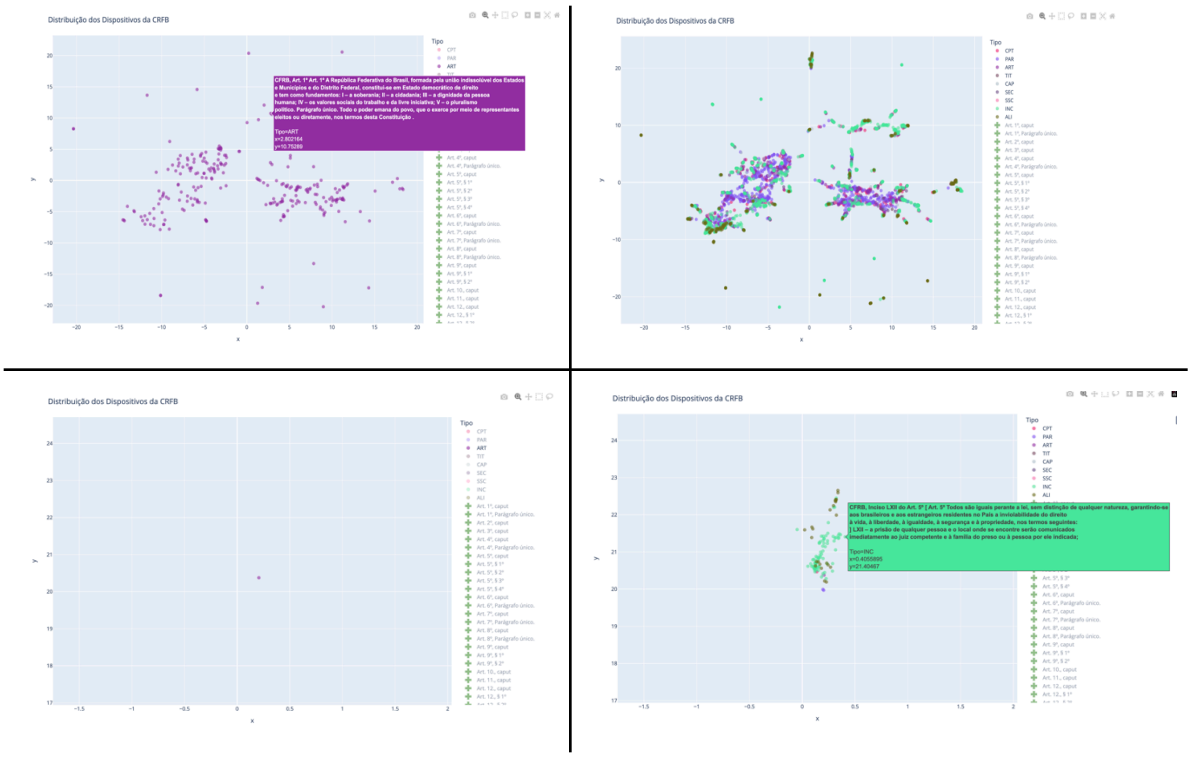
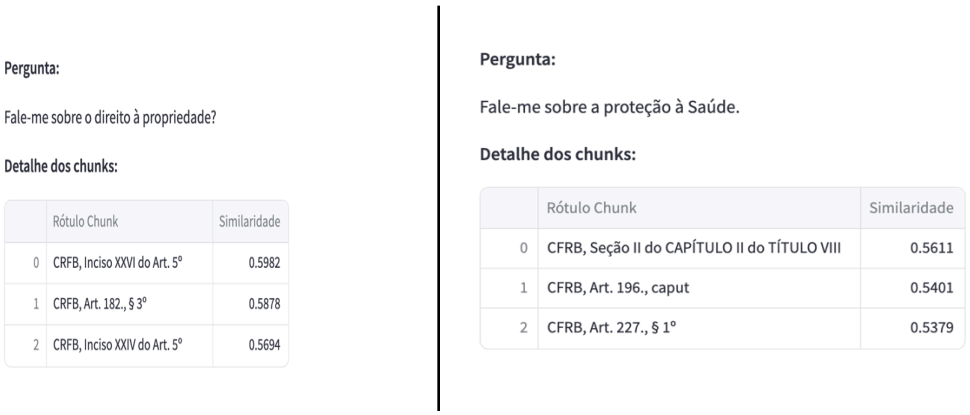


Figure 4. Multi-level Embedding Visualization of the Brazilian Federal Constitution

When comparing the third and fourth quadrants of Figure 4, it becomes evident that the fourth quadrant, in addition to possessing the vector representing Article 5 in its entirety, also contains dozens of other vectors representing more specific aspects of this important article that deals with fundamental rights. This allows the RAG strategy to offer more relevant excerpts in a more concise and, consequently, more economical manner. By utilizing language models through API, we pay for the quantity of input and output tokens. For example, when we ask "Tell me about the right to property", the similarity of the vectors identified 2 subsections of Article 5 and one paragraph of Article 182, as can be observed on the left side of Figure 5.

If the user formulates a more comprehensive question that encompasses topics that are part of several articles within the same grouping, the system is expected to return chunks of article groupings. This is what occurs in the example on the right side of Figure 5, where the user requests "Tell me about Health protection". Note that the first chunk returned was from Section II of Chapter II of Title VIII, which specifically deals with Health. This section has 5 articles (Articles 196 to 200).



Pergunta:

Fale-me sobre o direito à propriedade?

Detalhe dos chunks:

	Rótulo Chunk	Similaridade
0	CRFB, Inciso XXVI do Art. 5º	0.5982
1	CRFB, Art. 182., § 3º	0.5878
2	CRFB, Inciso XXIV do Art. 5º	0.5694

Pergunta:

Fale-me sobre a proteção à Saúde.

Detalhe dos chunks:

	Rótulo Chunk	Similaridade
0	CFRB, Seção II do CAPÍTULO II do TÍTULO VIII	0.5611
1	CFRB, Art. 196., caput	0.5401
2	CFRB, Art. 227., § 1º	0.5379

Figure 5. Most similar chunks for two questions

The flexibility of question embeddings presents a notable advancement in making legal searches more accessible. It allows users to query in simple language or even in different languages, bridging the gap between legal professionals and those without specialized knowledge. This versatility is particularly useful when dealing with fundamental rights, as it facilitates broader engagement and understanding.

The flexibility of question embeddings significantly enhances legal search by permitting queries in plain language or in various languages, narrowing the divide between legal practitioners and the general public. A legal expert might pose a question in technical terms such as 'Where does power emanate from?' (*'De onde emana o poder?'*) or, in Italian, *'Da dove viene il potere?'*. Alternatively, a layperson asking *'De onde vem a autoridade desse povo que manda?*' (Where does the authority of those who command come from?) will receive an equally accurate response. This capability illustrates the system's semantic depth—it prioritizes the collective meaning over individual words.

A BRAZILIAN BROADCASTING REGULATORY STATUTES EXPERIMENT

In order to evaluate our approach within a specific domain of Brazilian regulation, we preprocessed the embeddings of the main text from the Consolidation Decree GM/MCOM 1 of June 2, 2023 (MCOM-GM). This document encompasses 540 articles, organized into various hierarchical structures, totaling approximately 2,800 normative provisions. We further enriched this framework by incorporating embeddings from the Brazilian Federal Constitution (CRFB), the Civil Code (CC), and the Code of Civil Procedure (CPC), collectively adding about 12,000 additional normative provisions standardized within the proposed hierarchical structure. By integrating multiple normative bodies, we simulate a more realistic scenario where a legal query must be analyzed in the context of all related legislation. This comprehensive approach ensures a thorough examination of legal demands, reflecting the interconnected nature of legal systems.

In an initial experiment, we posed the question: “What should the radiation diagram of the antenna used by a Community Broadcasting Service station look like?” (*Como deve ser o diagrama de irradiação da antena utilizada por estação do Serviço de Radiodifusão Comunitária?*) Figure 6 illustrates the chunks obtained from the semantic search. These chunks were utilized to generate a natural language response: “The antenna irradiation diagram used by a Community Broadcasting Service station must be omnidirectional, as established in Art. 331 of the GM/MCOM Consolidation Decree No. 1/2023” (*O diagrama de irradiação da antena utilizada por estação do Serviço de Radiodifusão Comunitária deve ser omnidirecional, conforme estabelecido no Art. 331 da Portaria de Consolidação GM/MCOM nº 1/2023.*) Below is the literal content of the articles mentioned in the chunks in the original language. Notably, the final response text closely mirrors the wording of Art. 331:

> Art. 331. O diagrama de irradiação da antena utilizada por estação do Serviço de Radiodifusão Comunitária deverá ser omnidirecional. (Origem: PRT GM/MCOM 4.334/2015, art. 80, caput)
>
> Art. 327. A emissão deverá ter as seguintes características técnicas: (Origem: PRT GM/MCOM 4.334/2015, art. 76, caput)
>
> I – designação: monofônica: 180KF3EGN estereofônica: 256KF8EHF; (Origem: PRT GM/MCOM 4.334/2015, art. 76, I)

II – polarização: a polarização da onda eletromagnética emitida pela antena poderá ser linear (horizontal ou vertical), circular ou elíptica; (Origem: PRT GM/MCOM 4.334/2015, art. 76, II)

III – tolerância de frequência: a frequência central da estação de Radiodifusão Comunitária não poderá variar mais que ±2000 Hz de seu valor nominal; e (Origem: PRT GM/MCOM 4.334/2015, art. 76, III)

IV – espúrios de radiofrequência: qualquer emissão presente em frequências afastadas de 120 a 240 kHz, inclusive, da frequência da portadora deverá estar pelo menos 25 dB abaixo do nível da portadora sem modulação; as emissões em frequências afastadas de mais de 240 kHz até 600 kHz, inclusive, da frequência da portadora deverão estar pelo menos 35 dB abaixo do nível da portadora sem modulação; as emissões em frequências afastadas de mais de 600 kHz da frequência da portadora deverão estar pelo menos (73 + P) dB (P=potência de operação do transmissor, em dBk) abaixo do nível da portadora sem modulação.(Origem: PRT GM/MCOM 4.334/2015, art. 76, IV)

	Rótulo Chunk	Similaridade
0	MCOM-GM, Art. 331., caput	0.8778
1	MCOM-GM, Art. 331.	0.8424
2	MCOM-GM, Art. 327., caput	0.7526

Figure 6. Chunks for “What should the radiation diagram of the antenna used by a Community Broadcasting Service station look like?”

In a second case, we presented the following request: “Tell me about the license for station operation” (*Fale-me sobre a licença para funcionamento da estação*). Figure 7 displays the resulting chunks. Given the open-ended nature of this question, the entire content of Part III of the Decree was utilized, encompassing 216 articles with hundreds of additional provisions, spanning from articles 176 to 391. However, since the text of Article 379 (Art. 379) is more specific, it was presented first, followed by another specific article within its scope, which was also highlighted. By prioritizing the most relevant and specific articles, we ensured a comprehensive and precise response. This approach demonstrates the system’s capability to navigate extensive legal texts and efficiently extract pertinent information, thereby providing detailed insights into the licensing requirements for station operation.

	Rótulo Chunk	Similaridade
0	MCOM-GM, Art. 379., caput	0.7203
1	MCOM-GM, CAPÍTULO IV do TÍTULO III do LIVRO V da PARTE III	0.7189
2	MCOM-GM, Art. 203., caput	0.7084

Figure 7. Chunks for “Tell me about the license for station operation.”

The response to the second query consisted of five paragraphs. We present here only selected excerpts in the original language. By selectively presenting key excerpts, we highlight the most pertinent information from the comprehensive response, showcasing the system’s ability to generate detailed and contextually accurate answers for complex legal queries.

A licença para funcionamento da estação é um documento crucial no contexto da radiodifusão no Brasil, conforme estabelecido na Portaria de Consolidação GM/MCOM nº 1/2023. Este documento é emitido pelo

Ministério das Comunicações após a deliberação pelo Congresso Nacional e a expedição de um Decreto Legislativo, conforme descrito no art. 324. A licença tem um prazo de vigência de dez anos.

Para obter essa licença, são necessários [....]

Além disso, o art. 326 estabelece que o prazo para o início efetivo da execução do Serviço de Radiodifusão Comunitária é de seis meses a contar [...].

É importante destacar que a operação da estação em novo local só é permitida após a emissão de uma nova licença com as informações atualizadas, conforme o art. 379. Além disso, [...].

Em resumo, a licença para funcionamento da estação é um documento fundamental que formaliza a autorização para operação de uma estação de radiodifusão, detalhando todas as informações necessárias para sua operação legal e regulamentada dentro do território brasileiro.

CONCLUSION

By implementing this multi-layered approach and investigating further development avenues, we aim to create a comprehensive and flexible system for legal information retrieval, enabling more efficient and accurate access to legal knowledge.

The proposed a multi-layered embedding-based retrieval approach offers a promising avenue for advancing legal information retrieval. By capturing the inherent structure and diverse granularities of legal knowledge, this method facilitates more accurate, efficient, and adaptable information access, ultimately empowering legal professionals and researchers in their pursuit of legal knowledge.

Additional considerations for development:

- Inter-Article Relationships: Investigate methods to represent relationships between articles, such as cross-references and dependencies. This could involve creating embeddings for pairs or groups of articles, capturing how they interact and inform each other.
- Temporal Dimension: Consider incorporating a temporal dimension into the embeddings, reflecting the evolution of legal concepts and amendments over time. This could involve creating embeddings for different versions of the legal text or using time-aware embedding models.
- Vector Dimensions: During our proof-of-concept tests involving the Constitution, the Civil Code, the Civil Procedure Code and the Consolidation Decree, we utilized 256 dimensions, which yielded excellent empirical results. Given that our model is capable of supporting up to 3072 dimensions, future investigations should assess whether increasing the number of dimensions can enhance performance further. This exploration will help determine if the additional computational cost is justified by potentially improved accuracy and contextual sensitivity in the retrieval of complex legal documents, thus optimizing the trade-off between resource expenditure and retrieval efficacy.
- Future work may evaluate the use of more dimensions, compare different language models both for the preprocessing of embeddings and for the final generation of the response in natural language, and expand the legal database of norms.

REFERENCES

1. Ben Ayed, A. and Biskri, I. (2020). On the Relevance of Query Expansion Using Parallel Corpora and Word Embeddings to Boost Text Document Retrieval Precision. *Computer Science and Information Technology*, 9, 1-7.
2. Chersoni, E., Santus, E., Huang, C-R., Lenci, A. (2021). Decoding Word Embeddings with Brain-Based Semantic Features'. *Computational Linguistics*, 47, 1-36.
3. Mimouni, N., Nazarenko, A., Paul, È., Salotti, S. (2014). Towards Graph-based and Semantic Search in Legal Information Access Systems. *International Conference on Legal Knowledge and Information Systems*.
4. Ryle, G. (1933). About. *Analysis*, 1(2), 10-12.
5. Saravanan, M., Ravindran, B., Raman, S. (2009). Improving legal information retrieval using an ontological framework. *Artificial Intelligence and Law*, 17, 101-124.
6. Yu, M., Dredze, M. (2015). Learning Composition Models for Phrase Embeddings. *Transactions of the Association for Computational Linguistics*, 3, 227-242.