# HyLEX-RAG: Bridging the Lexical-Semantic Gap in Regulatory Compliance Retrieval

**Márcia Rodrigues Campos, Glauco Vitor Pedrosa**

[1]Graduate Program in Applied Computing (PPCA)
University of Brasilia (UnB) – Brasilia – DF – Brazil

marcitarcampos@gmail.com ,glauco.pedrosa@unb.br

***Abstract.*** *Regulatory information retrieval requires balancing semantic understanding with exact lexical precision. Pure semantic retrieval captures contextual meaning but suffers from normative ambiguity, while lexical approaches fail under paraphrases and synonym variations. This paper proposes HyLEX-RAG, a hybrid retrieval approach that combines hypothetical 5W1H question indexing with weighted lexical-semantic fusion for regulatory question answering. Experiments conducted on BACEN regulatory norms show that HyLEX-RAG achieved the best overall results, reaching 66.7% faithfulness, 70.0% answer rate, and 0% hallucination on out-of-scope questions, achieving better overall results than the evaluated baselines ($p < 0.05$). The results demonstrate that combining semantic and lexical retrieval enhances the reliability of RAG systems for compliance and auditing applications.*

## 1. Introduction

The continuous growth in the volume of regulatory documents in the financial sector has significantly increased the challenges associated with the accurate retrieval of normative information [Antoniou et al. 2022, Ashley 2017]. In the context of the Central Bank of Brazil (BACEN), resolutions, circulars, normative instructions, and other regulatory acts constitute a document collection characterized by technical legal language, dense normative content, and frequent cross-references. In this scenario, auditing, compliance, and regulatory supervision activities directly depend on the ability to quickly locate specific information, including legal obligations, regulatory exceptions, deadlines, operational procedures, and responsible entities [Petrov and Stocken 2022].

Retrieval-Augmented Generation (RAG) approaches have recently become widely adopted in question-answering systems over large document collections [Lewis et al. 2020]. In general, RAG systems combine information retrieval mechanisms with Large Language Models (LLMs), enabling the generation of answers grounded in documents external to the model itself. This integration allows generated answers to be grounded on retrieved documents rather than relying exclusively on parametric knowledge.

Most modern RAG approaches rely on semantic retrieval based on vector embeddings [Karpukhin et al. 2020, Reimers and Gurevych 2019]. However, in the regulatory domain, queries frequently involve resolution numbers, specific articles, dates, institutional acronyms, and highly specialized legal terminology. Consequently, even small semantic ambiguities may lead to the retrieval of incorrect passages, directly compromising the reliability of the generated answers. In regulatory tasks, a semantically

plausible but normatively incorrect answer may represent significant risks for auditing and compliance processes.

Furthermore, there is a relevant linguistic asymmetry between the way users formulate queries and the way regulatory documents are written. While queries are generally expressed in task-oriented natural language, operational problems, or practical doubts, regulatory texts are written using formal, procedural, and legal language. This discrepancy substantially reduces the effectiveness of traditional retrieval strategies based exclusively on direct textual similarity.

Motivated by these limitations, this work proposes *HyLEX-RAG* (*Hypothetical and Lexical Retrieval-Augmented Generation*), a hybrid retrieval architecture for regulatory consultation and auditing. The proposed approach combines semantic retrieval based on hypothetical questions structured according to the 5W1H model with BM25 lexical retrieval applied directly to the original normative chunks. Unlike purely embedding-based or purely lexical approaches, the method performs a weighted fusion between semantic and lexical scores, leveraging contextual similarity while preserving normative precision.

The central hypothesis of this work is that regulatory retrieval requires both semantic contextualization and exact lexical matching, since purely semantic similarity is insufficient for preserving normative precision in compliance-oriented question answering tasks. The goal of the proposed approach is not to replace lexical retrieval, but to preserve its normative precision while extending semantic robustness under paraphrased and operational queries.

To evaluate the proposed approach, a corpus containing BACEN (Central Bank of Brazil) regulations was constructed, subsequently segmented into textual chunks and enriched with hypothetical 5W1H questions automatically generated by LLMs. In addition, a benchmark composed of complex questions developed within the context of financial regulatory auditing was created. Different retrieval strategies were experimentally compared, including traditional RAG vector-based retrieval and the exclusive use of BM25-based retrieval.

The main contributions of this work are:

- the demonstration that hybrid lexical-semantic retrieval is more reliable than purely semantic retrieval for compliance-oriented regulatory question answering;
- an indexing strategy based on automatically generated 5W1H hypothetical questions;
- a weighted fusion mechanism between semantic and lexical scores for regulatory retrieval;
- an experimental evaluation over BACEN regulations using complex regulatory auditing questions.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 describes the proposed method. Section 4 details the conducted experiments and obtained results. Finally, Section 5 presents the conclusions and future work.

## 2. Related Work

Information retrieval in question-answering systems has evolved significantly with the emergence of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) architectures. However, applications in regulatory domains still face challenges related to normative precision, traceability, and semantic alignment between queries and documents.

### 2.1. Retrieval-Augmented Generation (RAG)

The RAG paradigm, introduced by [Lewis et al. 2020], combines document retrieval with LLM-based text generation by using external content as contextual support for response generation. This strategy mitigates limitations of static parametric knowledge and increases factual grounding.

Most RAG architectures rely on dense embedding-based retrieval [Karpukhin et al. 2020], enabling the identification of semantic relationships even without direct lexical matching. Models such as BERT [Devlin et al. 2019] and Sentence-BERT [Reimers and Gurevych 2019] have become widely adopted in this context. Nevertheless, purely vector-based approaches still present limitations in specialized scenarios that require high terminological and normative precision.

### 2.2. Lexical Retrieval and Hybrid Systems

Lexical methods such as BM25 [Robertson and Zaragoza 2009] remain highly relevant in legal and regulatory domains due to the need for exact matching of technical terms, legal articles, and normative references. Despite their robustness, these methods struggle with natural language queries and semantic reformulations [Manning et al. 2008].

To overcome these limitations, hybrid systems have been proposed by combining lexical and semantic retrieval. Studies such as [Lin et al. 2021] and results from the BEIR benchmark [Thakur et al. 2021] demonstrated that hybrid approaches frequently outperform exclusively dense or lexical methods.

### 2.3. Hypothetical Expansion and Semantic Retrieval

Recent research has investigated hypothetical expansion strategies to reduce the semantic gap between queries and documents using LLM-generated synthetic representations. HyDE (*Hypothetical Document Embeddings*) [Gao et al. 2023] generates hypothetical documents from user queries before vector retrieval, achieving significant improvements in *zero-shot* retrieval scenarios.

Recently, the work of [Vake et al. 2025] proposed HyPE-RAG, which uses hypothetical questions associated with documents as a semantic alignment mechanism. Unlike HyDE, this approach employs synthetic questions to represent multiple informational perspectives of the same document, increasing the semantic coverage of the vector index.

Despite recent advances in hypothetical expansion methods such as HyDE and HyPE-RAG, most approaches still assume that semantic similarity is sufficient for retrieval quality. This assumption becomes problematic in regulatory environments, where semantically plausible retrieval may still violate normative correctness. Existing literature has largely overlooked the interaction between semantic enrichment and exact lexical preservation in compliance-oriented retrieval tasks.

### 2.4. Information Retrieval in Legal and Regulatory Domains

The application of artificial intelligence to legal and regulatory documents has grown substantially in recent years. According to [Ashley 2017], intelligent legal systems strongly depend on accurate retrieval of norms, precedents, and legal references.

However, regulatory documents present characteristics such as highly formal language, strong contextual dependence, and high terminological density, which hinder traditional semantic retrieval approaches. Furthermore, semantically plausible but factually incorrect responses may introduce significant risks in auditing, compliance, and regulatory decision-making processes [Kokina et al. 2025].

Although recent advances have explored RAG in specialized domains, there is still a lack of hybrid approaches specifically designed for financial regulatory documents.

### 2.5. Positioning of This Work

While prior hypothetical-expansion methods reduce the query-document semantic gap, they share a common assumption: that semantic similarity alone is sufficient for retrieval quality. HyDE [Gao et al. 2023] generates hypothetical documents from queries before vector retrieval, and HyPE-RAG [Vake et al. 2025] associates hypothetical questions with documents to enrich the semantic index. In both cases, retrieval ultimately reduces to comparing embeddings, which is precisely where regulatory retrieval breaks down: a semantically plausible passage may still be normatively incorrect, missing the exact article, resolution number, acronym, or deadline that determines compliance.

Our approach differs in two ways. First, instead of free-form hypothetical questions, we structure question generation around the 5W1H model (who, what, when, where, why, how), so that each chunk is indexed along well-defined informational dimensions, involved entities, obligations, temporality, normative location, justification, and procedure rather than arbitrary paraphrases. This produces a more systematic semantic coverage of normative content. Second, and centrally, we do not rely on semantic enrichment alone: hypothetical-question embeddings are fused with BM25 lexical scores computed directly over the original normative chunks, through a weighted combination that explicitly preserves exact terminological matching.

To the best of our knowledge, the interaction between LLM-based semantic enrichment and exact lexical preservation has been largely overlooked in compliance-oriented retrieval. The present work targets this gap directly, proposing a hybrid strategy designed for regulatory scenarios that demand both semantic robustness and normative precision, and empirically demonstrating over BACEN regulations that this combination outperforms purely semantic, purely lexical, and prior hypothetical-expansion baselines.

## 3. Proposed Method

This work proposes *HyLEX-RAG* (*Hypothetical and Lexical Retrieval-Augmented Generation*), a hybrid information retrieval approach designed for querying and auditing regulatory documents from the Central Bank of Brazil (BACEN). The approach combines semantic retrieval based on hypothetical questions (5W1H) with lexical retrieval (BM25), aiming to reduce the linguistic asymmetry between natural language queries and the technical vocabulary present in regulatory texts.

Figure 1 presents an overview of the proposed approach. The method is divided into two main stages: (i) construction of the hybrid index and (ii) weighted hybrid retrieval for answer generation.

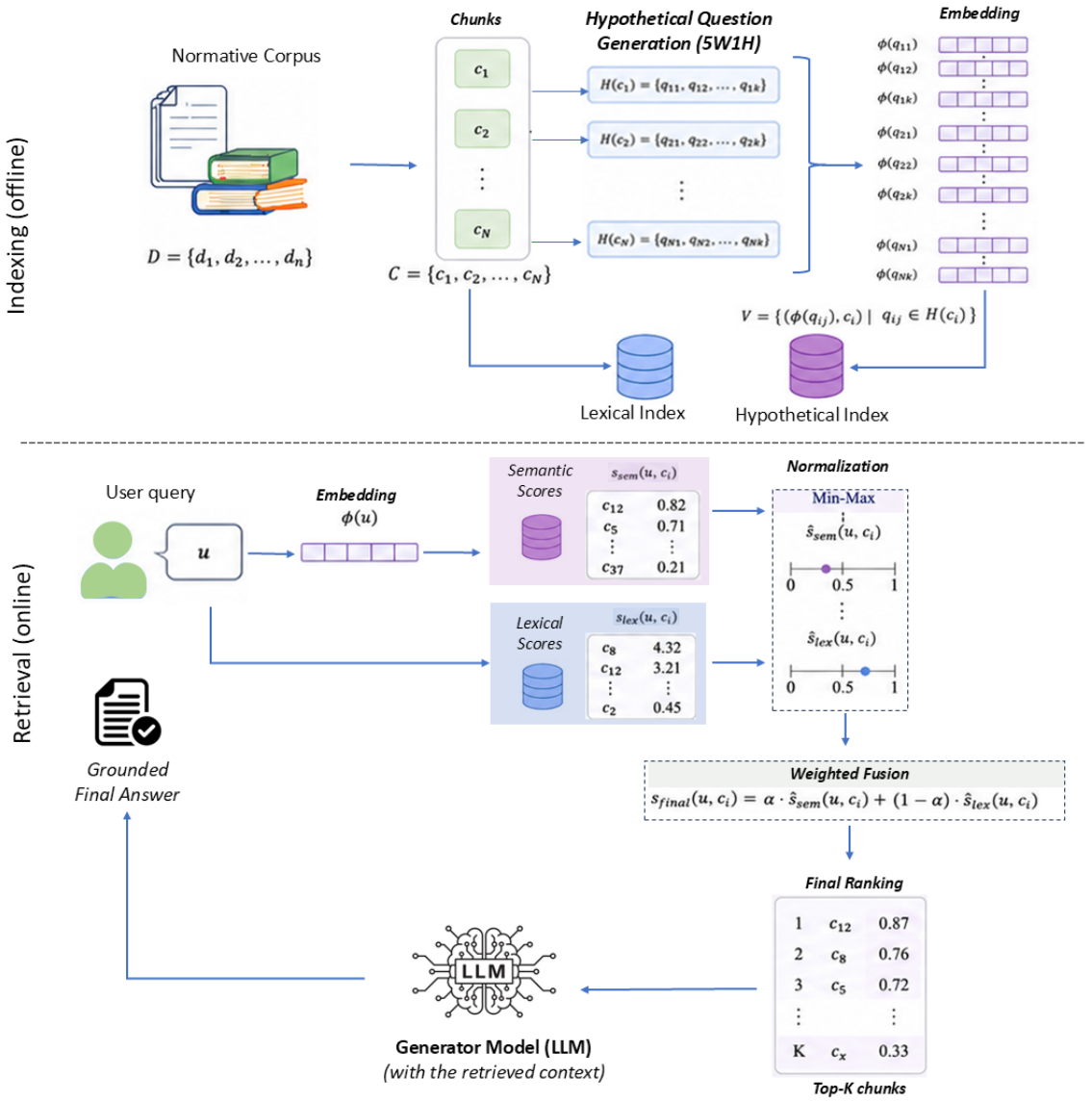



**Figure 1. Proposed HyLEX-RAG approach for hybrid semantic-lexical retrieval and grounded answer generation over normative corpus.**

### 3.1. Hypothetical Index Construction

Consider a normative corpus $D = \{d_1, d_2, \ldots, d_n\}$, segmented into a set of textual chunks $C = \{c_1, c_2, \ldots, c_N\}$. For each chunk $c_i \in C$, a set of $k$ hypothetical questions is automatically generated: $H(c_i) = \{q_{i1}, q_{i2}, \ldots, q_{ik}\}$. In this work, we adopted $k = 6$, using the 5W1H structure, i.e., $H(c_i) = \{q_{\text{who}}, q_{\text{what}}, q_{\text{when}}, q_{\text{where}}, q_{\text{why}}, q_{\text{how}}\}$

The use of the 5W1H structure aims to expand the semantic coverage of each chunk, enabling the representation of different informational perspectives of the normative content, including involved entities, obligations, temporality, normative location, regulatory justifications, and operational procedures.

The hypothetical questions are generated by a Large Language Model (LLM) and subsequently converted into vector embeddings through a semantic encoding function: $\phi : q_{ij} \rightarrow R^m$, where $\phi(q_{ij})$ represents the semantic vector associated with the hypothetical question.

The vector index of hypothetical questions is then defined as: $V = \{(\phi(q_{ij}), c_i) \mid q_{ij} \in H(c_i)\}$. Each embedding remains associated with its corresponding original chunk, allowing different hypothetical questions to point to the same normative passage.

In parallel, the original chunks are indexed using a lexical mechanism, such as BM25, forming a traditional textual index capable of capturing exact matches of technical terms, resolution numbers, dates, acronyms, and normative references.

### 3.2. Weighted Hybrid Retrieval

Given a user query $u$, the system simultaneously executes two retrieval mechanisms: one semantic and one lexical.

In the semantic stage, the query is converted into an embedding and compared against the indexed hypothetical questions:

$$s_{\text{sem}}(u, c_i) = \max_{q_{ij} \in H(c_i)} \text{sim}(\phi(u), \phi(q_{ij}))$$

where $\text{sim}(\cdot)$ represents cosine similarity between vectors.

The maximum operator is used because multiple hypothetical questions may represent different semantic aspects of the same chunk. Thus, the highest similarity obtained between the query and the questions associated with the chunk is considered.

In parallel, the lexical score is computed using BM25 directly over the original chunks:

$$s_{\text{lex}}(u, c_i) = \text{BM25}(u, c_i)$$

The scores produced by the semantic and lexical mechanisms operate on different scales. Therefore, both are normalized using Min-Max normalization:

$$\hat{s}(u, c_i) = \frac{s(u, c_i) - s_{\min}}{s_{\max} - s_{\min}}$$

resulting in the normalized scores $\hat{s}_{\text{sem}}$ and $\hat{s}_{\text{lex}}$.

The core stage of HyLEX-RAG consists of the weighted fusion of the two retrieval mechanisms:

$$s_{\text{final}}(u, c_i) = \alpha \cdot \hat{s}_{\text{sem}}(u, c_i) + (1 - \alpha) \cdot \hat{s}_{\text{lex}}(u, c_i)$$

where $\alpha \in [0, 1]$ controls the relative contribution between semantic and lexical retrieval. Higher values of $\alpha$ favor semantic similarity, whereas lower values prioritize exact lexical matching.

After the fusion stage, the chunks are ranked in descending order of relevance according to the final score:

$$R(u) = \text{TopK}_{c_i \in C} \left( s_{\text{final}}(u, c_i) \right)$$

The top-$K$ most relevant chunks are then used as context in the Retrieval-Augmented Generation (RAG) stage.

## 4. Experimental Results

### 4.1. Experimental Setup

The experiments were conducted using a corpus composed of 20 regulations from the Central Bank of Brazil (BACEN), including resolutions, circulars, and normative instructions. The documents were segmented using a sliding window of 800 characters with an overlap of 250 characters, resulting in a total of 2,265 normative chunks. The chunk size was empirically selected to preserve complete normative clauses while minimizing context fragmentation.

The generative model used for both hypothetical question generation and answer generation was Gemma 3 12B-It. Vector retrieval employed embeddings generated by the `all-MiniLM-L6-v2` model, while vector storage was implemented using FAISS. For lexical retrieval, BM25 was applied over the original normative chunks.

The experimental evaluation considered four distinct retrieval approaches:

- **Traditional RAG:** retrieval based exclusively on vector similarity between the query and the chunks;
- **HyPE-RAG:** retrieval based on embeddings of hypothetical questions;
- **BM25:** purely lexical retrieval over the normative chunks;
- **HyLEX-RAG:** the approach proposed in this work, based on weighted fusion between semantic and lexical scores.

The experiments were conducted using a benchmark composed of 50 complex questions developed within the context of financial regulatory auditing, including 10 lexical questions, 30 semantic questions, and 10 control questions. The queries involved normative articles, regulatory obligations, deadlines, exceptions, institutional acronyms, and cross-references.

### 4.2. Evaluation Metrics

The evaluation employed both heuristic metrics and metrics based on the RAGAS framework [Es et al. 2024]. The heuristic metrics included:

- **F1 Tokens**: measures textual overlap between the generated answer and the expected answer;
- **Response Rate**: percentage of queries effectively answered by the system;
- **Numerical Precision**: measures the correctness of numbers, dates, and normative references present in the generated responses;

Additionally, automatic metrics from the RAGAS framework were employed:

- **Faithfulness**: measures the fidelity of the answer to the retrieved context;
- **Context Precision**: evaluates the relevance of the retrieved context;
- **Context Recall**: measures contextual coverage with respect to the expected answer.

The RAGAS evaluation used Gemini 2.5 Flash as the judge model and `text-embedding-004` for auxiliary embeddings.

### 4.3. Overall Comparison Between Approaches

Table 1 presents the overall results of the evaluated retrieval approaches. Traditional vector-based RAG achieved the lowest performance, while HyPE-RAG provided moderate improvements through hypothetical-question indexing. BM25 showed strong normative precision, highlighting the importance of exact lexical matching in regulatory domains. HyLEX-RAG achieved the best overall results, with the highest F1 Tokens score (24.4%) and response rate (70.0%), indicating that combining semantic and lexical retrieval improves both contextual coverage and regulatory precision.

Although token-level F1 scores remain relatively low, this behavior is expected in regulatory QA scenarios, where multiple semantically valid answers may differ lexically while still preserving normative correctness.

**Table 1. Overall comparison between evaluated retrieval strategies**

| Method | F1 Tokens (%) | Response Rate (%) | Numerical Precision (%) |
|---|---|---|---|
| Traditional RAG | 9.5 | 32.5 | 18.8 |
| HyPE-RAG | 11.2 | 32.5 | 21.3 |
| BM25 | 21.6 | 65.0 | 30.5 |
| HyLEX-RAG | **24.4** | **70.0** | **42.3** |

Table 2 presents the results obtained by the different approaches using the RAGAS metrics. Traditional vector-based RAG achieved the lowest values for *faithfulness*, *context precision*, and *context recall*, indicating limitations in retrieving relevant context and generating answers faithful to the document content. HyLEX-RAG achieved the highest faithfulness (66.7%), matching BM25 in context precision (57.50%). For context recall, BM25 obtained a marginally higher value (43.75% vs. 42.50%), although the difference is negligible. Overall, HyLEX-RAG delivered the strongest balance between faithful generation and contextual relevance.

**Table 2. RAGAS metrics results**

| Approach | Faithfulness (%) | Context Precision (%) | Context Recall (%) |
|---|---|---|---|
| Traditional RAG | 28.24 | 27.50 | 15.00 |
| HyPE-RAG | 31.35 | 27.50 | 17.50 |
| BM25 | 61.04 | **57.50** | **43.75** |
| HyLEX-RAG | **66.7** | **57.50** | 42.50 |

The good performance of BM25 over purely semantic retrieval reinforces that regulatory retrieval differs substantially from open-domain QA, since exact references

to legal articles, acronyms, dates, and procedural obligations frequently dominate retrieval relevance. This behavior is particularly important in compliance and auditing environments, where confidently incorrect answers may introduce operational, legal, and regulatory risks.

### 4.4. Quantitative and Qualitative Analysis

In addition to the overall evaluation, the experimental benchmark was divided into two query categories: predominantly lexical questions and predominantly semantic questions. Table 3 presents the results obtained for each category.

**Table 3. Results by query category (F1-Tokens / Faithfulness, %)**

| | Lexical | | Semantic | |
|---|---|---|---|---|
| **Method** | F1 (%) | Faith (%) | F1 (%) | Faith (%) |
| Traditional RAG | 5.7 | 25.0 | 10.7 | 29.3 |
| HyPE-RAG | 13.2 | 40.0 | 10.5 | 28.5 |
| BM25 | **17.1** | 60.0 | 23.1 | 61.4 |
| HyLEX-RAG | 16.1 | **60.1** | **26.5** | **65.3** |

The results suggest that HyLEX-RAG performed consistently across both query categories. For lexical queries, BM25 achieved the highest F1-Tokens score (17.1%), likely benefiting from exact term matching, while HyLEX-RAG obtained a similar F1 score (16.1%) and comparable faithfulness (60.1% vs. 60.0%). For semantic queries, HyLEX-RAG achieved the highest F1-Tokens score (26.5%) and faithfulness (65.3%), indicating that the combination of semantic and lexical retrieval may contribute to improved performance when handling semantically expressed information needs.

Table 4 illustrates a representative example from the evaluation set. Traditional RAG and BM25 retrieved only partial information related to the query, while HyPE-RAG recovered a more relevant regulatory passage but did not fully address the transparency requirement. In contrast, HyLEX-RAG generated the answer most aligned with the ground truth, correctly identifying the obligation to provide daily information on constituted receivables and additional details required by the regulation.

### 4.5. Sensitivity Analysis of Parameter $\alpha$

A sensitivity analysis was also conducted for parameter $\alpha$, responsible for balancing semantic and lexical retrieval. Table 5 presents the results obtained for different values of $\alpha$.

The results suggest that extreme values of $\alpha$ are associated with lower performance. Higher values favor semantic retrieval, which may struggle with regulatory references such as acronyms, dates, and resolution numbers, whereas lower values increase reliance on exact lexical matching and reduce the benefits of semantic generalization. This behavior reinforces the complementary nature of lexical and semantic retrieval in regulatory question answering.

The best results were obtained with $\alpha = 0.5$, indicating that lexical retrieval and purely semantic retrieval act as complementary mechanisms in the financial regulatory domain.

**Table 4. Comparison between the ground truth and the answers generated by the evaluated methods. Examples are presented in Portuguese because the evaluated regulatory corpus is composed of original BACEN documents.**

**Question:** Quais são os critérios de transparência exigidos pelo regulador para instituições que realizam operações de desconto de recebíveis?

*Ground-Truth:* A instituição financeira deve disponibilizar ao usuário final recebedor informações sobre o valor de recebíveis constituídos a serem liquidados em cada dia.

| **Method** | **Retrieved Answer** |
|---|---|
| Traditional RAG | I - especificar os recebíveis de arranjo de pagamento constituídos e a constituir que sejam objeto da operação, no caso de operação de desconto. |
| HyPE-RAG | A identificação detalhada dos recebíveis objeto de desconto ou dados em garantia deve ser realizada pelo sistema de registro em que eles estejam registrados, conforme regras definidas no contrato de crédito, obedecida a regulamentação específica do Banco Central do Brasil. |
| BM25 | I - especificar os recebíveis de arranjo de pagamento constituídos e a constituir que sejam objeto da operação, no caso de operação de desconto; |
| HyLEX-RAG | A instituição financeira autorizada a visualizar os recebíveis de arranjo de pagamento do usuário final recebedor deve disponibilizar a esse usuário, no mínimo, as informações sobre o valor de recebíveis constituídos a serem liquidados em cada dia, desagregado por: I - valor de recebíveis bloqueados no dia que já tenham sido objeto de negociação ou entregues em garantia; e II - valor de recebíveis livres para negociação no dia. |

### 4.6. Statistical Test

Given the limited benchmark size and the non-Gaussian distribution of retrieval scores, a non-parametric paired Wilcoxon test was adopted. Table 6 presents the paired Wilcoxon test results obtained over the same evaluation queries. The results indicate statistically significant improvements of HyLEX-RAG over all evaluated baselines, as all p-values remained below the significance threshold of 0.05.

## 5. Conclusion

This work presented HyLEX-RAG, a hybrid retrieval approach for question answering and auditing over financial regulatory documents from the Central Bank of Brazil (BACEN). The method combines semantic retrieval based on hypothetical 5W1H questions with BM25 lexical retrieval to address the semantic-lexical mismatch between natural language queries and specialized regulatory texts.

The results showed that semantic retrieval alone is insufficient for regulatory scenarios requiring high normative precision, particularly when dealing with legal references, acronyms, dates, and procedural obligations. Lexical retrieval preserves normative precision but is less effective with paraphrased or reformulated queries. By combining both approaches, HyLEX-RAG achieved the best overall balance, obtaining the highest faithfulness and response rate, strong context precision, and robustness against out-of-scope hallucinations. Statistical analysis also confirmed significant improvements over the semantic and lexical baselines.

**Table 5. Sensitivity analysis of parameter $\alpha$ in HyLEX-RAG**

| $\alpha$ | F1 Tokens (%) | Numerical Precision (%) | Faithfulness (%) |
|---|---|---|---|
| 0.00 | 21.6 | 40.5 | 58.6 |
| 0.25 | 21.7 | 40.5 | 54.6 |
| 0.50 | **24.4** | **42.3** | **66.7** |
| 0.75 | 14.9 | 27.7 | 50.0 |
| 1.00 | 11.2 | 21.3 | 32.1 |

**Table 6. Paired Wilcoxon test comparing HyLEX-RAG with the other methods**

| Comparison | p-value |
|---|---|
| HyLEX-RAG vs Traditional RAG | $< 0.001$ |
| HyLEX-RAG vs HyPE-RAG | 0.003 |
| HyLEX-RAG vs BM25 | 0.007 |

The findings demonstrate that reliable RAG systems for regulatory compliance and auditing need to combine semantic understanding with exact normative matching. However, the study was limited by a small benchmark of 50 BACEN-related questions, a single experimental execution, and the use of LLM-based evaluation metrics, which may introduce bias.

Future work includes larger benchmarks, multiple experimental runs, human validation, adaptive fusion strategies, neural re-ranking, and domain-specific embeddings to further improve the reliability and generalizability of regulatory RAG systems.

## Declaration on the Use of Generative AI

The authors declare that generative Artificial Intelligence (AI) tools (GPT-5.6 Luna) were used exclusively to support language revision, grammar improvement, and text refinement during the preparation of this manuscript. All scientific content, methodological decisions, analyses, interpretations, and conclusions were conceived, verified, and approved by the authors, who assume full responsibility for the content of this work.

## References


Antoniou, G., Atkinson, K., Baryannis, G., Batsakis, S., Caro, L. D., Governatori, G., Robaldo, L., Siragusa, G., and Tachmazidis, I. (2022). Explainable reasoning with legal big data: A layered framework. *Journal of Applied Logics — IfCoLog Journal of Logics and their Applications*, 9(4):1155–1170.

Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pages 4171–4186.

Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th*

*Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gao, L., Ma, X., Lin, J., and Callan, J. (2023). Precise zero-shot dense retrieval without relevance labels. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1762–1777.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Kokina, J., Blanchette, S., Davenport, T. H., and Pachamanova, D. (2025). Challenges and opportunities for artificial intelligence in auditing: Evidence from the field. *International Journal of Accounting Information Systems*, 56:100734.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., K"uttler, H., Lewis, M., Yih, W.-t., Rockt"aschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Lin, J., Ma, X., Lin, S.-C., Yang, J.-H., Pradeep, R., and Nogueira, R. (2021). Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2356–2362.

Manning, C. D., Raghavan, P., and Sch"utze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Petrov, E. and Stocken, P. C. (2022). Auditor specialization and information spillovers. *The Accounting Review*, 97(7):401–428.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Thakur, N., Reimers, N., R"uckl'e, A., Srivastava, A., and Gurevych, I. (2021). Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks Track*.

Vake, D. et al. (2025). Bridging the question-answer gap in retrieval-augmented generation: Hypothetical prompt embeddings. *IEEE Access*.