Enhancing Regulatory Compliance Through Automated Retrieval, Reranking, and Answer Generation

Kübranur Umar*
TOBB University of Economics and Technology
kumar@etu.edu.tr

Hakan Doğan*
TOBB University of Economics and Technology
hakandogan@etu.edu.tr

Onur Özcan*
TOBB University of Economics and Technology
onurozcan@etu.edu.tr

İsmail Karakaya
HAVELSAN Inc.
ikarakaya@havelsan.com.tr

Alper Karamanlıoğlu
HAVELSAN Inc.
alperk@havelsan.com.tr

Berkan Demirel
HAVELSAN Inc.
bdemirel@havelsan.com.tr

Abstract

This paper explains a Retrieval-Augmented Generation (RAG) pipeline that optimizes regularity compliance using a combination of embedding models (*i.e.* bge-m3, jina-embeddings-v3, e5-large-v2) with reranker (*i.e.* bge-reranker-v2-m3). To efficiently process long context passages, we introduce *context aware chunking* method. By using the RePASS metric, we ensure comprehensive coverage of obligations and minimizes contradictions, thereby setting a new benchmark for RAG-based regulatory compliance systems. The experimental results show that our best configuration achieves a score of **0.79** in Recall@10 and **0.66** in MAP@10 with LLaMA-3.1-8B model for answer generation.

1 Introduction

Regulatory documents are critical components for many industries including finance, healthcare and insurance, to comply with standards and laws. These documents are characterized by complex legal terminology, hierarchical structures, and frequent updates. Therefore, this creates difficulties for interpretation and implementation. These incompatibilities lead to negative outcomes such as significant financial penalties, loss of reputation, and operational disruptions.

The complexity of regulatory documents to put forward the necessity for advanced systems capable of efficient information retrieval and synthesis. Retrieval-Augmented Generation (RAG) systems offers a promising solution for retrieval mechanism and answer generation.

Previous research in Regulatory Natural Language Processing (RegNLP) discovered the potential of machine learning for automating regulatory compliance, but some difficulties still exist:

1. High-precision retrieval of relevant passages from large regulatory corpora is challenging.

2. Ranking and synthesizing retrieved passages to ensure completeness and scope of obligation is another challenge.

3. Efficient processing of long contextual queries where relevant information may span multiple sections of a document, is another major challenge.

In this study, to address these challenges, we propose an optimized RAG pipeline for advanced-level ranking and improved generative performance, using a context-aware chunking strategy combined with *"bge-m3 + hybrid search"*, and *"bge-reranker-v2"*. Our contributions are as follows:

- Introducing a chunk-based approach for processing long regulatory contexts effectively.

- Evaluation of multiple retrieval and re-ranking models for regulatory QA tasks using the RePASS metric.

2 Related Work

The significant progress on RegNLP mostly about complexities of regulatory texts. The structured data extraction is focused by the previous studies. In this context, Lau et al. (2005) focus on XML-based frameworks in order to extract information from accessibility regulations. Also, Kiyavitskaya et al. (2008) propose the Cerno framework to focus on automation of rights and obligation extraction

*These authors contributed equally to this work

Proceedings of the 31st International Conference on Computational Linguistics, pages 91–96
January 19–24, 2025.

from legal texts. However, these works are insufficient in scalability and adaptability.

Thanks to the advent of deep learning, considerable improvements achieved in RegNLP. In this context, Chalkidis et al. (2018) introduce a hierarchical BiLSTM model in order to extract obligations from legal contracts, and this study outperforms the previous methods that relies on manual features. In addition, Nair et al. (2018) implement deep learning pipelines in the work of annotating global trade regulations. This method enables enhanced compliance workflows in the field of RegNLP. Similar to these works, Chalkidis et al. (2021) leverage BERT-based models to handle complex queries in EU/UK legislative texts. This method shows how transformer architecture is effective in processing long documents. Abualhaija et al. (2022) extend this method with BERT for automated question-answering (QA) systems targeting GDPR-related texts. Thanks to this work, a considerable success has achieved in passage retrieval tasks.

Gokhan et al. (2024) provide a baseline framework for regulatory QA tasks by introducing the ObliQA dataset[1] curated to address multi-passage queries. This dataset is as collection of over 27,000 QA pairs derived from Abu Dhabi Global Markets[2] regulations. Additionally, this study introduces Regulatory Passage Answer Stability Score (RePASS), a novel evaluation metric designed to measure the accuracy and consistency of generated answers in regulatory contexts. They combine sparse and dense retrieval methods (*e.g.*, BM25 and BGE models) with a generative approach to synthesize answers from retrieved passages. Despite the contribution of this work, they challenged in handling complex or lengthy queries, and the generative model exhibited limitations in contextual comprehension and obligation coverage.

RegNLP applications accelerated by the recent advancements in synthetic data generation. An example of these upgrades is QA dataset for roundtrip validation in Alberti et al. (2019). Also, Maatouk et al. (2023) propose zero-shot learning method for neural passage retrieval.

The integration of retrieval and generative models enables advanced QA methodologies in RAG systems. Lewis et al. (2020) formalize RAG as a framework that enriches generative models with retrieved knowledge. The retrieval efficiency and response quality is improved by Self-RAG (Asai et al. (2023)) and PipeRAG (Jiang et al. (2024)) systems having limited adaptation to regulatory texts.

Our study addresses challenges in retrieval precision and generative accuracy for regulatory QA. We introduce a robust RAG pipeline incorporating hybrid retrieval using dense models, advance re-ranking and context-aware chunking to manage long regulatory documents. This system achieves a Recall@10 of **0.79** and MAP@10 of **0.66**, establishing a new standard for regulatory question answering.

3 Methodology

The long passages in regulatory documents affect the performance of generative models, since processing extended contexts efficiently is a challenge for these models. In order to handle this challenge, we segment long passages into smaller chunks, then filter and re-rank to optimize the input for generative models. In the next sections, we describe the proposed methodology by explaining the retrieval pipeline, long-context processing techniques, and answer generation system. The demonstration of our pipeline is shown in Figure 1.

3.1 Retrieval Pipeline

Combination of retrieval and re-ranking models in the retrieval pipeline maximizes the recall and precision. This system ensures that the most relevant passages are prioritized for downstream processing.

3.1.1 Passage Retrieval

In the first retrieval stage, we experiment with multiple dense retrieval models, including **bge-m3**, **jina-embeddings-v3**, and **e5-large**. According to the obtained results, the **bge-m3** model outperforms other models and achieves Recall@10 of **0.74**. The results are detailed in Table 1. This results show that the model is suitable for regulatory texts, thanks to its ability to effectively capture semantic nuances.

In order to improve retrieval performance, a hybrid search mechanism is implemented combining dense (vector-based) and lexical retrieval methods. We achieve the best recall by tuning the hybrid search parameter to **0.3**. Top-50 passages are retrieved for each query that serves as input for the re-ranker model.

[1] https://github.com/RegNLP/ObliQADataset
[2] https://www.adgm.com/

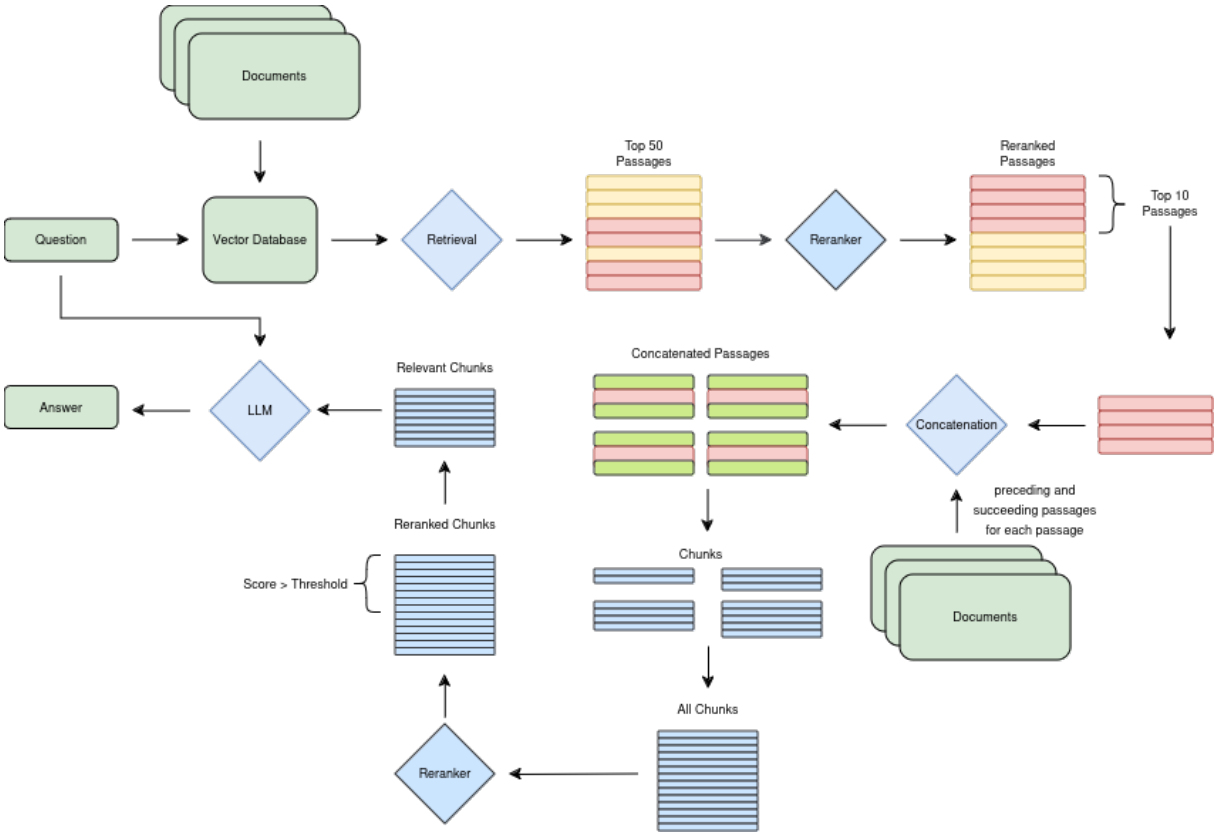

Figure 1: Pipeline Diagram: Retrieval, Reranking, and Generative Model Integration.

	w=0.5		w=0.3	
Model	**Recall@10**	**MAP@10**	**Recall@10**	**MAP@10**
jina-embeddings-v3	0.71	0.60	0.76	0.62
e5-large-v2	0.69	0.59	0.76	0.62
bge-m3	0.74	0.61	0.76	0.62
bge-m3+bge-reranker-v2-m3	0.77	0.65	**0.79**	**0.66**

Table 1: Retrieval and Reranking Performance.

3.1.2 Passage Re-ranking

We evaluate the **bge-reranker-v3-m3** model to improve the ranking of retrieved passages. We achieve the highest performance, with a **0.79** Recall@10, by the combination of **bge-m3** and **bge-reranker-v2**, when the hybrid search hyperparameter is set to **0.3**. In contrast, when the hybrid parameter is set to **0.5**, the Recall@10 value is **0.77**. These results indicate the importance of hyperparameter optimization in achieving high retrieval performance.

The re-ranker assigns points to retrieved passages and prioritizes the scope of obligation and relevance. The top 10 highest-scoring passages, according to their scores, are selected for further processing, which significantly improves the quality of inputs to the generative model.

3.2 Long Context Processing

Regulatory queries usually require synthesizing information span over more than one section. To handle this difficulty, we use a strategy that contains context expansion, chunking, and chunk filtering and re-ranking. This strategy is detailed below:

1. **Context Expansion**: Retrieved passages are enriched by their preceding and succeeding sections. This additional context improves the system's ability to address cross-referenced information and capture narrow regulatory obligations.

2. **Chunking**: Expanded passages are divided into smaller chunks in accordance with the input limitations of generative models *e.g.* **LLaMA-3.1-8B-Instruct**, with a maximum

Method	Es	Cs	OCs	RePASs
bge-m3+bge-reranker-v2-m3+LLaMA-3.1-8B-Instruct	0.39	0.30	0.12	0.41

Table 2: RePAS Scores

length of 1024 tokens per chunk and a stride value of 100 tokens. This operation proposes efficient processing while protecting critical regulatory information.

3. **Chunk Filtering and Re-ranking**: Each 1024-token segment provided as input to the model is processed through the **bge-reranker-v2-m3** model to enhance its performance and efficacy. The re-ranking process prioritizes chunks that are both relevant and contextually comprehensive, resulting in improved generative performance. The reranker model filters out less relevant chunks. This process reduces noise in the input data and ensures the generative model focuses on the most critical regulatory information. This streamlined input allows the model to generate more precise, contextually aligned, and reliable outputs, ultimately enhancing the accuracy and utility of the system for regulatory question-answering tasks.

   Re-ranker scores chunks for relevance and contextual completeness. Chunks that exceed a certain threshold are given as input to the model. In order to select high-quality chunks, a threshold score of 0.7 is applied. The threshold is reduced incrementally by 0.1 until at least one chunk meets the criteria because of some cases where all chunks score below this threshold. This process ensures that there are always input data for the generative model to process. By this way, the pipeline ensures that the generative model processes only the most relevant and high-quality parts, reducing noise and improving response accuracy.

3.3 Answer Generation

Filtered parts are given to the **LLaMA-3.1-8B-Instruct** model along with the query, and responses are generated. The generative model is used with one-shot prompt showed in Figure A.1 of Appendix A. The model is configured with the parameter `max_new_tokens` set to **512**. This allows the model to produce short but comprehensive answers. By presenting only the most relevant parts to the generative model, we ensure that its output is context-appropriate and compliant with regulatory requirements.

To evaluate the system performance, we use the **ObliQA dataset** which is introduced by Gokhan et al. (2024). This dataset consists of 27.869 QA pairs collected from financial regulations and provides a robust benchmark for regulatory QA systems. The evaluation is performed using the **RePASS metric** proposed by Gokhan et al. (2024). This metric evaluates obligation coverage, contradiction avoidance, and overall entailment. By using these tools, we ensure that it is compatible with the standards set for regulatory QA. The evaluation results, including RePASS values, are presented in detail in Table 2.

This methodology systematically addresses regulatory QA challenges by combining advanced retrieval techniques, efficient long-context processing capabilities, and careful filtering of inputs for generative models. By optimizing each stage of the pipeline, we demonstrate significant improvements in both retrieval accuracy and response quality.

4 Conclusion

This paper proposes RAG pipeline for regulatory compliance tasks by integrating hybrid search, advanced re-ranking and context-aware chunking strategies. RAG performance improved precision and recall significantly by using **bge-m3** model for hybrid search and **bge-reranker-v2** model for re-ranker. According to the experimental results, we achieve a score of **0.79** at Recall@10 and **0.66** at MAP@10. Introducing the chunk-based processing approach enhanced the **LLaMA-3.1-8B-Instruct** model's generative capabilities and enabled more effective processing of long-context regulatory documents.

References

Sallam Abualhaija, Chetan Arora, Amin Sleimi, and Lionel C Briand. 2022. Automated question answering for improved understanding of compliance requirements: A multi-document study. In *2022 IEEE 30th international requirements engineering conference (RE)*, pages 39–50. IEEE.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin,

and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. *arXiv preprint arXiv:1906.05416*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2018. Obligation and prohibition extraction using hierarchical rnns. *arXiv preprint arXiv:1805.03871*.

Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou, and Prodromos Malakasiotis. 2021. Regulatory compliance through doc2doc information retrieval: A case study in eu/uk legislation where text similarity has limitations. *arXiv preprint arXiv:2101.10726*.

Tuba Gokhan, Kexin Wang, Iryna Gurevych, and Ted Briscoe. 2024. Regnlp in action: Facilitating compliance through automated information retrieval and answer generation. *arXiv preprint arXiv:2409.05677*.

Wenqi Jiang, Shuai Zhang, Boran Han, Jie Wang, Bernie Wang, and Tim Kraska. 2024. Piperag: Fast retrieval-augmented generation via algorithm-system co-design. *arXiv preprint arXiv:2403.05676*.

Nadzeya Kiyavitskaya, Nicola Zeni, Travis D Breaux, Annie I Antón, James R Cordy, Luisa Mich, and John Mylopoulos. 2008. Automating the extraction of rights and obligations for regulatory compliance. In *Conceptual Modeling-ER 2008: 27th International Conference on Conceptual Modeling, Barcelona, Spain, October 20-24, 2008. Proceedings 27*, pages 154–168. Springer.

Gloria T Lau, Kincho H Law, and Gio Wiederhold. 2005. Legal information retrieval and application to e-rulemaking. In *Proceedings of the 10th international conference on Artificial intelligence and law*, pages 146–154.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Ali Maatouk, Fadhel Ayed, Nicola Piovesan, Antonio De Domenico, Merouane Debbah, and Zhi-Quan Luo. 2023. Teleqna: A benchmark dataset to assess large language models telecommunications knowledge. *arXiv preprint arXiv:2310.15051*.

Rahul Nair, Killian Levacher, and Martin Stephenson. 2018. Towards automated extraction of business constraints from unstructured regulatory text. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 157–160.

A Appendix

One-Shot Prompt

You are a knowledgeable assistant specialized in regulatory documents. Your task is to generate precise answers to questions based on the given context.

Example:

Question: *What are the specific approval requirements and criteria set forth by the ADGM for a Shari'a Supervisory Board's endorsement of a Public Fund's Constitution and Prospectus?*
Context :

- *The Fund Manager of a Domestic Fund that is a Public Fund must ensure that its Fund's Constitution and Prospectus are, and remain, approved by the Fund's Shari'a Supervisory Board.*
- *The Fund Manager of an Exempt Fund must ensure that the Fund's Constitution and Prospectus are, and remain, approved by the Fund Manager's Shari'a Supervisory Board.*

Answer: *The ADGM requires that the Constitution and Prospectus of a Public Fund be approved by a Shari'a Supervisory Board. For Domestic Public Funds, this approval must come from the Fund's Shari'a Supervisory Board, whereas for Exempt Funds, it is the Fund Manager's Shari'a Supervisory Board that must provide and maintain this approval.*

Now, answer the following question based on the given context:

Question: Question Text
Context : Relative Chunks
Answer:

Figure A.1: One-Shot Prompt Representation.