NUST Nova at RIRAG 2025: A Hybrid Framework for Regulatory Information Retrieval and Question Answering

Mariam Babar Khan, Huma Ameer, Seemab Latif, and Mehwish Fatima*
School of Electrical Engineering and Computer Science (SEECS),
National University of Sciences and Technology (NUST),
Islamabad, Pakistan
[mariamk.bscs21seecs,hameer.msds20seecs,
seemab.latif,mehwish.fatima]@seecs.edu.pk

Abstract

NUST Nova participates in RIRAG Shared Task, addressing two critical challenges: Task 1 involves retrieving relevant subsections from regulatory documents based on user queries, while Task 2 focuses on generating concise, contextually accurate answers using the retrieved information. We propose a Hybrid Retrieval Framework that combines graph-based retrieval, vector-based methods, and keyword matching (BM25) to enhance relevance and precision in regulatory QA. Using score-based fusion and iterative refinement, the framework retrieves the top 10 relevant passages, which are then used by an LLM to generate accurate, context-aware answers. After empirical evaluation, we also conduct an error analysis to identify our framework's limitations.

1 Introduction

The Regulatory Information Retrieval and Answer Generation (RIRAG) shared task focuses on advancing Question Answering (QA) in regulatory compliance. Participants develop systems to retrieve relevant information and generate precise answers to complex compliance queries, addressing the critical need for interpreting specialized regulatory language in domains like legal research and policy analysis.

Traditional Information Retrieval (IR) methods like BM25 (Robertson et al., 1994) excel at keyword-based document retrieval but struggle with the nuanced, context-dependent language of regulatory texts (de Andrade and Becker, 2023; Yang et al., 2023). Vector-based retrieval, leveraging document embeddings, shows promising results but faces challenges with domain-specific terminology and maintaining relevance (Monir et al., 2024; Sarmah et al., 2024). Similarly, graph-based retrieval excels in regulatory contexts but suffers from scalability issues and handling ambiguous or incomplete data (Jain et al., 2023; Technology, 2015; Sarmah et al., 2024). These limitations underscore the need for hybrid approaches to enhance precision and scalability in regulatory text retrieval (Sarmah et al., 2024).

To address the challenges of regulatory QA, we propose a Hybrid Retrieval Framework with multi-method scoring to enhance passage retrieval precision. The framework combines three models: (1) Neo4j, which structures queries and passages into a graph for initial relevance extraction, (2) BM25 for keyword matching, and (3) FAISS for ranking passages based on semantic similarity. These models are fused through score-based fusion, refining results by combining BM25 and FAISS outputs with those from Neo4j. This hybrid approach ensures accurate retrieval of the top 10 passages. Finally, the Llama model generates context-aware, regulatory-compliant answers, effectively handling domain-specific terminology and complex relationships.

2 Hybrid Retrieval Framework

Our hybrid retrieval[1] system integrates knowledge-graph and vector-based methods, combining their strengths to enhance accuracy and relevance. A score fusion mechanism merges relevance scores, followed by re-ranking to produce a balanced, high-quality ranked list. An LLM processes the top-ranked passages for context-aware, regulatory-compliant answers. Figure 1 illustrates this integration for efficient results.

2.1 Embeddings Generation

We use LegalBERT (LB) (Chalkidis et al., 2020) to generate dense embeddings for regulatory information retrieval and answer generation. LB provides domain-specific knowledge critical for understanding complex legal content as it is pre-trained on a large corpus of legal and regulatory texts. These

*Corresponding author: mehwish.fatima@seecs.edu.pk

[1] https://github.com/MehwishFatimah/NUST-Nova.git

Proceedings of the 31st International Conference on Computational Linguistics, pages 73–78
January 19–24, 2025.

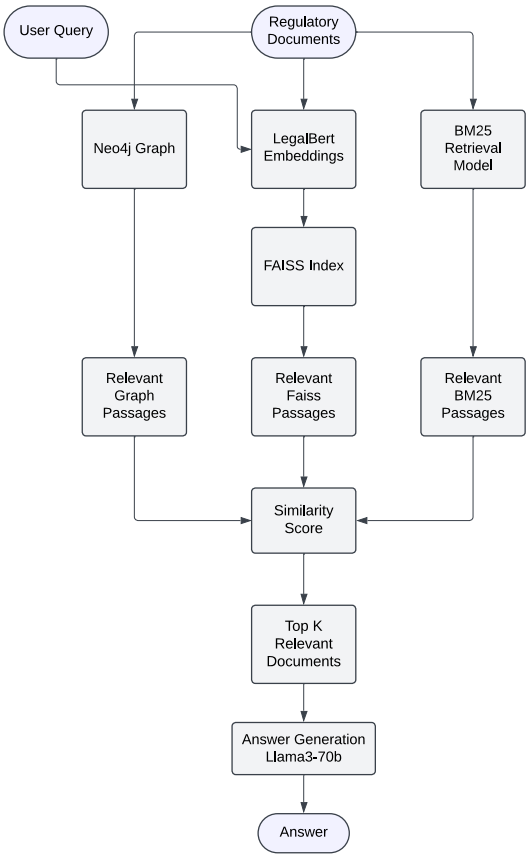

Figure 1: Hybrid Retrieval Framework: three retrieval models work simultaneously and then results are ranked before answer generation

embeddings, optimized for legal semantics, are stored for vector-based retrieval using FAISS(Douze et al., 2024) and enable efficient similarity computations.

2.2 Information Retrieval

For IR, we implement a hybrid approach that combines graph-based retrieval, vector-based retrieval, and traditional BM25.

2.2.1 Graph-Based Retrieval

We use Neo4j Graph Database (NEO4J) (Technology, 2015) to enhance the retrieval of relevant passages based on the structural relationships within the data. We query our NEO4J database for a specific question by retrieving passages that are connected to the query's node through direct relationships in the graph. Using LB, the query text is converted into a numerical vector representation. Relevance is determined by calculating the cosine similarity between query and passage embeddings, with the top-ranked passages returned based on their similarity scores.

2.2.2 Statistical Retrieval

The tokenized dataset is indexed using BM25 (Robertson et al., 1994), a well-established ranking function in information retrieval that evaluates lexical overlap between the query and document passages. BM25 assigns relevance scores based on factors like term frequency, inverse document frequency, and query term saturation, effectively ranking passages by relevance. It ranks document passages based on their relevance to the query, and the top-ranked passages are retrieved for further processing.

2.2.3 Vector-Based Retrieval

For the vector database-based approach, we first generate high-quality vector embeddings for regulatory data using LB. These embeddings are indexed with FAISS for retrieval, where dense query embeddings, also generated by LB, are compared to pre-computed passage embeddings in the FAISS index using cosine similarity. This process efficiently retrieves the top-ranked passages, ensuring semantically accurate and relevant results.

2.3 Fusion and Re-ranking

To combine the results of NEO4J and BM25+FAISS, we use a score-based fusion approach. Initially, passages are retrieved independently by each method, with scores assigned based on their respective retrieval techniques. For graph-based retrieval, similarity scores are computed using cosine similarity between the query and linked passages' embeddings. BM25, a probabilistic model, calculates document relevance based on term frequency and inverse document frequency. In contrast, vector-based retrieval derives scores through approximate nearest-neighbor searches in the embedding space. The results from the mentioned methods are merged, eliminating duplicates and retaining the higher similarity score for overlapping passages. The combined passages are then re-ranked by recalculating their relevance using cosine similarity between the query embedding and the passage embeddings, ensuring that the most relevant passages, as identified by all retrieval methods, are ranked highest.

2.4 Answer Generation

The last step involves generating responses using the Llama3-70b model (LLAMA3) (Dubey et al., 2024). Passages retrieved from the Hybrid Retrieval Framework are concatenated and provided

as context to LLAMA3, which generates a coherent and accurate response tailored to the user's query. LLAMA3 synthesizes information from the retrieved passages to produce precise and contextually rich output. By using its pre-trained knowledge and the input passages to generate responses, LLAMA3 maintains the nuance and formal tone required for regulatory language.

Prompt engineering is crucial in our pipeline, ensuring generated responses align with regulatory obligations and avoid contradictions. Clear instructions are provided to cover all key requirements, structure responses, and align with source sentences from retrieved passages, reducing hallucinations and maintaining factual consistency.

Although fine-tuning is not yet implemented, future iterations will focus on adapting LLAMA3 to regulatory documents, enhancing its understanding of domain-specific jargon, hierarchical clauses, and inter-references. This will improve the model's ability to generate precise, compliant answers.

By leveraging structured prompts and relevant passages, LLAMA3 minimizes hallucinations, focusing on the most relevant context for generating accurate, high-quality responses.

3 Experiments

3.1 Dataset

We use the given ObliQA dataset which includes three subsets: the train set contains 22,295 questions, the test set have 2,786 questions, and the development set includes 2,888 questions. We use the train and development sets for evaluating various models and the final evaluation is performed on the unseen test set provided by the organizers. A comprehensive overview of the shared task and dataset are presented to Appendix C for brevity.

3.2 Baselines and Our Model

We consider BM25 (Gokhan et al., 2024) as our baseline due to its good performance. The top passages are processed by GPT-4-TURBO-1106 (GPT-4) (OpenAI, 2023) with a relevance threshold of 0.7, using a tailored prompt to generate compliance-focused answers that integrate regulatory requirements. We evaluate our final model and a variation of it: (i) NEO4J+BM25+FAISS, and (ii) BM25+FAISS. For answer generation, we use LLAMA3 to generate contextually accurate and coherent responses based on the retrieved passages.

3.3 Evaluation Metrics

We follow the standard metrics by organizers (Gokhan et al., 2024): MAP@10 and RECALL@10 for passage retrieval, and RePaSs ($\mathbf{Re}$) for answer generation.

Model	RECALL@10	MAP@10
BM25	0.76	0.62
BM25+FAISS	0.58	0.29
NEO4J+BM25+FAISS	0.79	0.74
NEO4J+BM25+FAISS	0.39	0.23

Table 1: Performance Comparison of Retrieval Models. The last row presents the results from organizers.

4 Results

Table 1 presents the results of retrieval models for RECALL@10 and MAP@10. The proposed framework NEO4J+BM25+FAISS achieves the highest scores of RECALL@10 = 0.79 and MAP@10 = 0.74 by using Neo4j's graph structure for capturing structural relationships among documents, while BM25 and FAISS ensure precise term matching and semantic alignment. This demonstrates the efficacy of integrating diverse retrieval strategies to address the complexity of regulatory texts.

The BM25 model demonstrates strong performance with RECALL@10 = 0.76 and MAP@10 = 0.62, confirming its reliability in retrieving relevant passages in a regulatory context. Its focus on exact term matching makes it particularly effective for structured legal texts, though it is limited in handling complex semantic relationships.

The BM25 model performs well (RECALL@10 = 0.76, MAP@10 = 0.62), excelling in regulatory contexts with its focus on exact term matching but struggling with semantic complexity. The BM25+FAISS model underperforms (RECALL@10 = 0.58, MAP@10 = 0.29), as FAISS's semantic retrieval weakens precision, highlighting misalignment with BM25 in domain-specific tasks.

4.1 Answer Generation Metrics

Table 2 compares the performance of two baseline methods: BM25+GPT-4 passage-only (PO) and BM25+GPT-4 rank fusion (RF) against two hybrid approaches: NEO4J+BM25+FAISS+LLAMA3 and BM25+FAISS+LLAMA3. The baselines achieve high relevance scores ($\mathbf{E_S}$ = 0.77, 0.77) but decline in contextual accuracy ($\mathbf{C_S}$ = 0.24, 0.24) and open-ended query handling ($\mathbf{OC_S}$ = 0.22, 0.20), resulting in $\mathbf{Re}$ scores of 0.58 and 0.58. This high-

Models	$\mathbf{E_S}$	$\mathbf{C_S}$	$\mathbf{OC_S}$	Re
BM25(PO)+GPT-4	0.77	0.24	0.22	0.58
BM25(RF)+GPT-4	0.77	0.24	0.20	0.58
BM25+FAISS+LLAMA3	0.31	0.25	0.07	0.37
NEO4J+BM25+FAISS+LLAMA3	0.43	0.36	0.15	0.41
NEO4J+BM25+FAISS+LLAMA3	0.36	0.31	0.11	0.39

Table 2: Comparison of Answer Generation Performance. The last row presents the results from organizers.

lights the limitations of keyword-based retrieval for nuanced regulatory queries.

NEO4J+BM25+FAISS+LLAMA3 shows moderate performance ($\mathbf{E_S}$ = 0.43, $\mathbf{C_S}$ = 0.36, $\mathbf{OC_S}$ = 0.15, $\mathbf{Re}$ = 0.41). Its graph-based integration improves semantic retrieval but struggles with open-ended queries. BM25+FAISS+LLAMA3 underperforms, with low relevance ($\mathbf{E_S}$ = 0.31), moderate contextual accuracy ($\mathbf{C_S}$ = 0.25), and poor open-ended query handling ($\mathbf{OC_S}$ = 0.07), yielding a $\mathbf{Re}$ score of 0.37. This highlights that vector-based retrieval alone is inadequate for regulatory QA without structured graph-based methods.

These results show that baseline models excel in relevance but struggle with contextual accuracy and open-ended queries. Hybrid methods improve structured retrieval via graph-based techniques but require optimization to balance relevance and adaptability for regulatory QA.

4.2 Error Analysis

We conduct an in-depth error analysis on 446 unseen questions to identify Hybrid Retrieval Framework's limitations in Appendix A. For this purpose, we apply a multi-step approach to evaluate the performance and quality of the responses generated by the model.

4.2.1 Data Preprocessing

First, we process the dataset by categorizing the questions based on whether an answer was generated or not. We split questions with empty answers and those with generated answers into two groups. we then preprocess data by tokenizing and filtering out stopwords to ensure the format suitable for analysis.

4.2.2 Topic Modeling

To explore further, we apply topic modeling using Latent Dirichlet Allocation (LDA) to identify prevalent themes in both groups of questions. This allow us to analyze the distribution of topics within the questions with empty answers and with generated answers. We evaluate these results to get insights

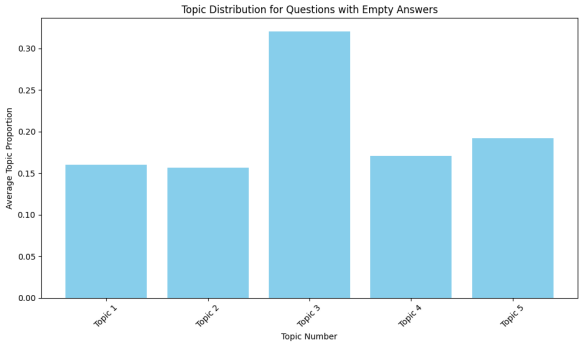

Figure 2: Topic Distribution of Questions with Empty Answers

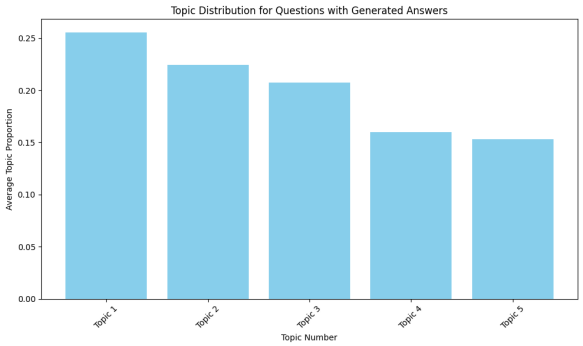

Figure 3: Topic Distribution of Questions with Generated Answers

about quality and relevance of the generated answers. LDA reveals distinct patterns in topic distributions illustrated in Figures 2 and 3. The details of these topics are presented in Appendix 4.2.2.

5 Conclusion

This work presents a hybrid framework combining vector-based, graph-based, and keyword-matching techniques to enhance regulatory information retrieval and answer generation. The approach significantly improves relevance and contextual accuracy, especially in handling domain-specific content. Preliminary results show improvements over baseline methods, with promising retrieval performance. However, answer generation results require refinement, highlighting the need for further enhancement. Future work includes exploring different LLMs or fine-tuning them for regulatory data and incorporating summarization techniques to optimize answer generation and extending graph-based retrieval to operate on entire documents rather than individual passages.

Acknowledgments

We sincerely thank the organizers and reviewers for their valuable contributions, constructive feedback, and support.

References

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school.

Leonardo de Andrade and Karin Becker. 2023. Bb25hlegalsum: Leveraging bm25 and bert-based clustering for the summarization of legal documents. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 255–263.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2023. Bayesian optimization based score fusion of linguistic approaches for improving legal document summarization. *Knowledge-Based Systems*, 264:110336.

Solmaz Seyed Monir, Irene Lau, Shubing Yang, and Dongfang Zhao. 2024. Vectorsearch: Enhancing document retrieval with semantic embeddings and optimized search. *arXiv preprint arXiv:2409.17383*.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1994. Okapi at trec-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Bhaskarjit Sarmah, Dhagash Mehta, Benika Hall, Rohan Rao, Sunil Patel, and Stefano Pasquali. 2024. Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 608–616.

Inc Technology. 2015. Neo4j, the world's leading graph database. *Neo4j Graph Database*.

Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2023. Empower large language model to perform better on industrial domain-specific question answering. *arXiv preprint arXiv:2305.11541*.

A Limitations

The graph-based approach in the framework faces scalability challenges, as managing and querying large, dynamic regulatory datasets can become resource-intensive, leading to slower retrieval times and higher computational costs. Additionally, the reliance on pre-trained models like BM25, FAISS, and Neo4j, while effective, limits adaptability to the nuanced language of the regulatory domain, affecting precision in handling domain-specific variations. Analysis of questions with empty answers reveals the system's strength in addressing specialized queries, with topics such as "adgm", "compliance", and "regulations" highlighting its focus on financial and regulatory concepts. However, for broader or less specific inquiries, the system struggles to maintain relevance, as indicated by topics like "customer" and "business". This highlights a gap in handling ambiguous or general questions, suggesting the need for enhanced contextual interpretation to improve performance across diverse query types.

B Training Considerations

Our framework avoids custom training, using pre-trained retrieval techniques for efficiency. This eliminates the need for resource-intensive model training while maintaining strong relevance for regulatory QA tasks.

C Task and Data

The RIRAG shared task comprises two key components: *Task 1: Regulatory Information Retrieval* focuses on retrieving relevant passages from complex, domain-specific regulatory documents in response to user queries, and *Task 2: Regulatory Answer Generation* involves producing concise, accurate answers based on the retrieved passages. Together, these tasks aim to advance the development of models that improve the accuracy and reliability of systems addressing complex regulatory queries.

The ObliQA dataset (Gokhan et al., 2024) advances Regulatory NLP (REGNLP) by providing 40 structured regulatory documents from Abu Dhabi Global Markets (ADGM), governing financial services in UAE free zones. With subsections, numbered clauses, and cross-references, it is well-suited for compliance applications. Converted to JSON

format, the dataset is validated using the DEBERTA-V3-XSMALL model (He et al., 2021) across three classes: Entailment, Contradiction, and Neutral.

D Error Analysis

Topic modeling on questions with empty answers revealed distinct themes. **Topic 1** encompassed terms like "risk", "person", "authorised", "adgm", and "management", reflecting a focus on risk and authorization processes in the ADGM context. **Topic 2** highlighted words such as "provide", "could", "specific", "risk", and "requirements", indicating queries related to precise regulatory risks and compliance criteria. **Topic 3** emphasized "virtual", "assets", "specific", "adgm", and "requirements", underscoring questions about virtual asset regulations. Similarly, **Topic 4** involved "could", "requirements", "guidance", "risk", and "adgm", pointing to inquiries about regulatory guidance. Lastly, **Topic 5** featured terms like "regulator", "person", "rule", "adgm", and "reporting", focusing on reporting standards and regulatory rules. These themes provide insights into gaps in the system's ability to generate answers and highlight areas for enhancement.

Topic modeling on questions with generated answers revealed five distinct themes. **Topic 1** was characterized by terms such as "compliance", "reporting", "virtual", "must", and "adgm", indicating a focus on regulatory compliance and mandatory reporting requirements. **Topic 2** featured terms like "provide", "information", "customer", "business", and "could", suggesting queries related to customer or business-specific information needs. **Topic 3** emphasized "financial", "risk", "must", "person", and "authorised", highlighting themes around financial risk and regulatory authorizations. **Topic 4** included terms such as "financial", "treatment", "standards", "per", and "could", reflecting inquiries about financial treatment and adherence to standards. Lastly, **Topic 5** was defined by terms like "adgm", "risk", "reporting", "person", and "regulations", focusing on risk management and regulatory reporting within the context of the Abu Dhabi Global Market (ADGM). These topics collectively provide insights into the nature of questions for which the system successfully generated answers.