Regulatory Question-Answering Using Generative AI

Devin Quinn[1]*, **Sumit Pai**[2]*, **Nirmala Pudota**[2]*, **Iman Yousfi**[1], **Sanmitra Bhattacharya**[1]
[1]Deloitte & Touche LLP, United States
[2]Deloitte & Touche Assurance & Enterprise Risk Services India Private Limited, India

Abstract

Although retrieval augmented generation (RAG) has proven to be an effective approach for creating question-answering systems on a corpus of documents, there is a need to improve the performance of these systems, especially in the regulatory domain where clear and accurate answers are required. This paper outlines the methodology used in our submission to the Regulatory Information Retrieval and Answer Generation (RIRAG) shared task at the Regulatory Natural Language Processing Workshop (RegNLP 2025). The goal is to improve document retrieval (Shared Task 1) and answer generation (Shared Task 2). Our pipeline is constructed as a two-step process for Shared Task 1. In the first step, we utilize a text-embedding-ada-002-based retriever, followed by a RankGPT-based re-ranker. The ranked results of Task 1 are then used to generate responses to user queries in Shared Task 2 through a prompt-based approach using GPT-4o[1]. For Shared Task 1, we achieved a recall rate of 75%, and with the prompts we developed, we were able to generate coherent answers for Shared Task 2.

1 Introduction

Regulations are official rules and directives established and maintained by authoritative bodies, such as government or regulatory agencies, to ensure compliance with legal standards. They are crucial for maintaining order, protecting public interests, and fostering fair practices across various industries. Due to the extensive range of regulations and the intricate nature of the language used in the regulatory content, comprehending these guidelines can be challenging for both the general public and regulatory professionals. Failure to adhere to regulations can result in legal and financial consequences, adversely impacting an organization's reputation and operations.

The RIRAG shared task (Gokhan et al., 2024) aims to improve the efficiency and accuracy of compliance-related tasks within the regulatory domain by encouraging the development of advanced Information Retrieval (IR) and answer generation techniques. When presented with a regulatory question, the main objective is to extract relevant passages from a vast collection of regulatory documents from Abu Dhabi Global Markets (ADGM)[2], which oversees financial services in the UAE's free economic zones. These extracted passages are then used to generate coherent and contextually accurate responses to the queries. The details of this dataset are described in Section 3.

In our submission, we address both shared tasks: Passage Retrieval (Subtask 1) and Answer Generation (Subtask 2). Our system design is presented in Section 4. Evaluation results on the development and test set is shown in Section 5. Finally, we conclude and discuss the next steps in Section 6.

2 Related Work

Early efforts in RegNLP concentrated on pattern matching, rule-based and semantic relation extraction methods. However, devising these patterns and rules can be quite difficult due to the complex nature and style of regulatory texts. Traditional information retrieval methods such as Best Matching-25 (BM25) and Term Frequency-Inverse Document Frequency (TF-IDF) have been widely used for regulatory information retrieval (Rosa et al., 2021)(Lau et al., 2003). But, these methods often struggle with shifts in word distribution and fail to adequately capture semantic similarity between words. Regulatory information retrieval using modern machine learning approaches (Ash and Chen, 2017; Tang et al., 2016; Collarana et al., 2018) such as word/document embeddings, Recurrent neural networks (RNN), and Long Short Term Memory

*These authors contributed equally to this work

[1]GPT-4o was selected based on performance as demonstrated on the HELM leaderboard

[2]https://www.adgm.com

Proceedings of the 31st International Conference on Computational Linguistics, pages 102–106
January 19–24, 2025.

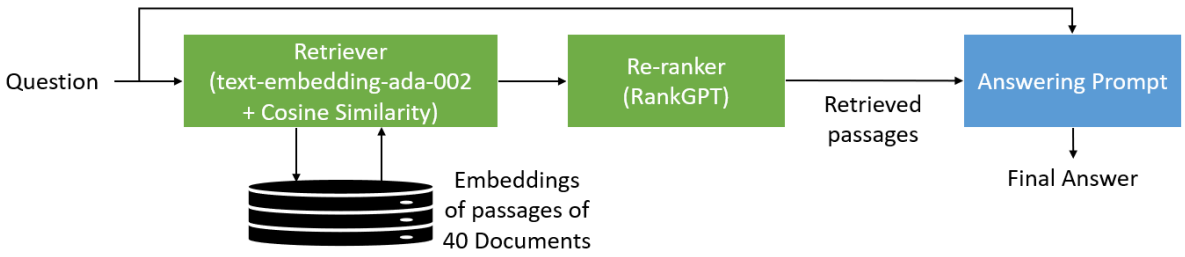

Figure 1: System Architecture

networks (LSTM) are good at modeling language and recognizing semantic similarities among various words and passages. However, they fall short in capturing long-range dependencies. Transformer-based approaches have demonstrated notable advancements in retrieval performance, as highlighted by several studies relevant to regulatory retrieval (Louis and Spanakis, 2021; Schumann et al., 2022) (Su et al., 2024). However, they still need significant amount of annotated datasets for fine-tuning.

With the advent of Large Language Models (LLMs), we can now accomplish a diverse range of NLP-related tasks without requiring task-specific fine-tuning. Ranking is one such area where these models particularly excel. RankGPT (Sun et al., 2023) proposes a list-wise ranking methodology (Ma et al., 2023; Pradeep et al., 2023) that addresses the issue of the LLM's context length by employing a sliding-window technique. Passages are ranked within each window, which then shifts incrementally to cover the entire list, ensuring overall re-ranking while staying within the LLM's context length limitation. For the task of answer-generation in regulatory domain, LSTM and transformer-based architectures have demonstrated significant results as highlighted by Collarana et al. 2018 (Zhong et al., 2020), but they require domain-specific labeled data for training. Recently, LLMs, through prompt engineering (Liu et al., 2021; Reynolds and McDonell, 2021) and RAG (Lewis et al., 2021) have shown promising results in generating coherent and grounded answers in the provided context.

In our paper, we use text-embedding-ada-002 from OpenAI for the retrieval followed by RankGPT for re-ranking, the details of which are elaborated in Section 4. The retrieved passages are then used as context within an engineered prompt to generate answers to the user query using GPT-4o.

3 Dataset

The Obligation-Based Question Answering (ObliQA) dataset (Gokhan et al., 2024), was specifically developed to support research in regulatory compliance. It includes question-answer pairs derived from passages in regulatory documents provided by the ADGM financial authority. These passages were selected individually or identified through topic-based clustering. Question-answer pairs were generated utilizing the GPT-4 model. To maintain precision and relevance, the generated questions were meticulously filtered for strong semantic alignment with the corresponding passages. Based on the number of passages used to generate the answers, the dataset was categorized into groups, where each group contains different combinations of 1 to 6 input passages per question. The data was subsequently divided into training, development, and test sets, containing 22k, 2.7k, and 2.7k samples, respectively.

4 System Design

The combined architecture of our system for both subtasks is shown in Figure 1 and detailed in the following sub-sections.

4.1 Subtask 1: Passage Retrieval

The components highlighted in green correspond to Subtask 1. We use text-embedding-ada-2 to embed the passages of ObliQA dataset and a standard vector database with a cosine similarity retriever. When a query is presented, we embed the query with the same embedding model and compute cosine similarity between the query embedding and the passage embeddings to retrieve the 30 most semantically similar passages. These initially retrieved passages are then input into RankGPT, which functions as a re-ranker as explained in Section 2, to reorder the passages and return its top 10. The final rankings reported for Subtask 1 were

derived from this process.

As we developed our approach, we experimented with several different embedding models/techniques for RAG. Due to resource and time constraints, we limit our exploration to a few relatively small models, but of various sizes (ranging from 100M params to 8b params) and embedding dimension sizes (ranging from 768 to 4096). We evaluate base and fine-tuned versions of all-mpnet-base-v2, legal-bert-base-uncased, and Qwen2.5-1.5B-Instruct on a small subset of test data, but ultimately choose text-embedding-ada-2 as it showed the best performance metrics on it.

4.2 Subtask 2: Answer Generation

We design and iteratively improve a prompt that optimizes RePaSs. This prompt incorporates the user query and the passages retrieved from Subtask 1, generating contextually grounded answers (the blue component in Figure 1). This prompt is outlined in Table 3 and is used with GPT-4o to generate the relevant answers.

5 Results

To evaluate the effectiveness of our methodology, we first establish a baseline retrieval using text-embedding-ada-2 for comparison which only returns the initial top-30 results without any re-ranking. For an initial qualitative analysis, we use a small subset (10%) of the test data. The baseline system achieved 70% recall@10, while re-ranking using RankGPT demonstrated a 5% improvement over this. An example of this improvement is shown in Table 4, where RankGPT successfully re-ranked a ground-truth reference passage that was initially not within the top 10, which the baseline retriever missed. This passage provided critical context for a comprehensive answer, which our method captured accurately, unlike the baseline. Encouraged by these initial qualitative results, we proceed to conduct evaluations on the full datasets.

The performance of our method on the full development and test sets for Subtask 1 is summarized in Table 1. Our approach gets a good recall and mean average precision (MAP). A group-wise and passage-wise analysis reveals that model performance diminishes as complexity rises. Specifically, recall@10 and MAP@10 scores are high for the retrieval of single passages; however, these metrics decline as the number of passages to be retrieved increases. This trend is consistent across different groups, indicating that the model's ability to effectively retrieve and rank passages declines as the quantity of relevant passages grows.

On Subtask 2, we evaluate the RePASs metric as described in Gokhan et al. 2024. Our prompts achieve a high entailment score (E_S), indicating that the answers are well-supported by the source passages. However, performance on obligation coverage (OC_S) is comparatively lower. The Overall Composite Score remains consistent across both datasets and is relatively good, as shown in Table 2 (comparable to the performance of the best models (Gokhan et al., 2024) on evaluation dataset).

6 Conclusion and Future Work

Our retrieval followed by re-ranking methodology demonstrates consistent and relatively good performance on both sets. However, as complexity increases, the effectiveness declines. We intend to investigate contrastive fine-tuning of retrievers to enhance retrieval capabilities and implement corrective-RAG for better contextual understanding, thereby delivering more relevant responses. Similarly, for answer generation, we observe that there is some degree of contradiction in answers compared to source passages as indicated by relatively high value for C_S. We aim to explore mechanisms to detect and resolve these contradictions and improve the obligation coverage (OC_S) by updating the answer generation prompt or potentially using a secondary prompt for refinement.

	Development		**Test**	
Subset	**R@10**	**M@10**	**R@10**	**M@10**
Full	75.7	60.3	75.3	59.7
G1	98.4	43.1	99.4	43.0
G2	71.0	23.1	72.7	25.8
G3	72.1	25.3	69.8	25.5
G4	60.1	23.9	58.7	23.1
G10	55.2	19.3	55.0	18.5
P1	84.0	32.8	83.9	33.0
P2	53.8	20.6	52.8	19.7
P3	38.8	14.5	35.2	15.3
P4	21.7	9.5	24.3	8.7
P5	36.7	19.0	26.7	12.0
P6	16.7	2.1	16.7	8.0

Table 1: Results of our approach on the development and test sets. Recall @10 and MAP@10 are reported on the Full dataset and on different groups and passage retrievals.

Type	E_S	C_S	OC_S	RePASs
Development	83.1	28.8	18.6	57.6
Test	82.7	28.1	19.4	58.0

Table 2: Performance of RePASs on Development and Test sets.

You are an editor of a regulatory magazine.
You are given a regulatory question and a set of passages as premise.
Your task is to understand the question and answer based on the provided passages.
———————————————
Here is the question: <question>
———————————————
Here are the passages: <passages>
———————————————
Your task is to carefully read the question and generate an answer based on the provided passages.

While generating the answer from the passages, you must strictly adhere to the following instructions:
1. The answer must not contain any statements that contradict the information in the passages.
2. The answer must encompass all the obligations present in the passages, ensuring that all critical regulatory obligations are reflected.
3. Every sentence in the answer must be substantiated by a corresponding sentence in the source passage.
4. Do not amalgamate information from multiple passages into a single sentence. Instead, create distinct sentences for each piece of information.
5. Provide a detailed answer, ensuring that all relevant information from the passages is included.

Generate an answer for the above question by strictly following these instructions, using the provided passages as the premise.
Output only the answer and nothing else. There is no need to justify or provide citations to the passages.
Simply generate the answer for the provided question based on the passages.
GENERATED ANSWER:

Table 3: Answer Generation Prompt for Subtask 2

Question:

With respect to Financial Instruments and Regulated Activities documentation under point (b),
could you please detail the level of granularity required in the records for each transaction or advice provided to a client?

(Original) Answer without re-ranking:

An authorised person must, pursuant to rule 6.8.2(b), make a record of the following in respect of each transaction:
the identity and account number of the client for whom the transaction was executed, or an indication that the transaction was for its own account;
the name of the counterparty;
the date and time where the transaction was executed;
the identity of the employee executing the transaction;
the investment, including the number of instruments or their value and price;
and whether the transaction was a purchase or a sale, long, short, buyer, seller or other relevant position.

(New) Answer with re-ranked passages:

An authorised person must, pursuant to rule 6.8.2(b), make a record of the following in respect of each transaction:
the identity and account number of the client for whom the transaction was executed, or an indication that the transaction was for its own account;
the name of the counterparty;
the date and time where the transaction was executed;
the identity of the employee executing the transaction;
the investment, including the number of instruments or their value and price;
and whether the transaction was a purchase or a sale, long, short, buyer, seller or other relevant position.
Additionally, an authorised person must, for a minimum of six years, maintain sufficient records in relation to each activity and function of the authorised person including any financial instruments provided to or regulated activities carried out for the benefit of a client and each advice or recommendation made to a client.

Ground truth context passage(s) missed before re-ranking:

An authorised person must, for a minimum of six years, maintain sufficient records in relation to each activity and function of the authorised person. these must include, where applicable, the following:
(a) any marketing material issued by, or on behalf of, the authorised person;
(b) any financial instruments provided to or regulated activities carried out for the benefit of a client and each advice or recommendation made to a client;
(c) documents regarding client classification under chapter 2;
(d) a record of each client agreement including any subsequent amendments to it as agreed with the client;
(e) records relating to the suitability assessment undertaken by the authorised person to demonstrate compliance with these rules;
(f) records to demonstrate compliance with the requirements relating to inducements, including any disclosure made to clients under that rule
and if any goods and services are received by the authorised person under a soft dollar agreement, the details relating to those agreements;
(g) financial promotions under schedule 2 of fsmr; and
(h) any other disclosures made to clients.

Table 4: Example of an input question and generated answers with/without re-ranking. In this example, RankGPT correctly re-ranked a reference passage missed by the baseline retriever, and subsequently the generated answer captures this necessary information while the original answer did not.

References

Elliott Ash and Daniel L Chen. 2017. Judge embeddings: Toward vector representations of legal belief. Technical report, Technical report.

Diego Collarana, Timm Heuss, Jens Lehmann, Ioanna Lytra, Gaurav Maheshwari, Rostislav Nedelchev, Thorsten Schmidt, and Priyansh Trivedi. 2018. A question answering system on regulatory documents. In *Legal knowledge and information systems*, pages 41–50. IOS Press.

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

Gloria T. Lau, Kincho H. Law, and Gio Wiederhold. 2003. Similarity analysis on government regulations. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, page 711–716, New York, NY, USA. Association for Computing Machinery.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Antoine Louis and Gerasimos Spanakis. 2021. A statutory article retrieval dataset in french. *arXiv preprint arXiv:2108.11792*.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm.

Guilherme Moraes Rosa, Ruan Chaves Rodrigues, Roberto Lotufo, and Rodrigo Nogueira. 2021. Yes, bm25 is a strong baseline for legal case retrieval.

Gerrit Schumann, Katharina Meyer, and Jorge Marx Gomez. 2022. Query-based retrieval of german regulatory documents for internal auditing purposes. In *2022 5th International Conference on Data Science and Information Technology (DSIT)*, pages 01–10. IEEE.

Weihang Su, Qingyao Ai, Yueyue Wu, Yixiao Ma, Haitao Li, Yiqun Liu, Zhijing Wu, and Min Zhang. 2024. Caseformer: Pre-training for legal case retrieval based on inter-case distinctions.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents.

Guoyu Tang, Honglei Guo, Zhili Guo, and Song Xu. 2016. Matching law cases and reference law provision with a neural attention model. *IBM China Research, Beijing*.

Botao Zhong, Wanlei He, Ziwei Huang, Peter E.D. Love, Junqing Tang, and Hanbin Luo. 2020. A building regulation question answering system: A deep learning methodology. *Advanced Engineering Informatics*, 46:101195.