RIRAG: Regulatory Information Retrieval and Answer Generation

Anonymous ACL submission

Abstract

Regulatory documents, issued by governmental regulatory bodies, establish rules, guidelines, and standards that organizations must adhere to for legal compliance. These documents, characterized by their length, complexity and frequent updates, are challenging to interpret, requiring significant allocation of time and expertise on the part of organizations to ensure ongoing compliance. Regulatory Natural Language Processing (RegNLP) is a multidisciplinary field aimed at simplifying access to and interpretation of regulatory rules and obligations. We develop a dataset based on generating question-passages pairs, where questions are automatically created and paired with relevant regulatory passages, facilitating the development of regulatory question-answering systems. We release the ObliQA dataset, containing 27,869 questions derived from the collection of Abu Dhabi Global Markets (ADGM) financial regulation documents, design a baseline Regulatory Information Retrieval and Answer Generation (RIRAG) system and evaluate it with RePASs, a novel evaluation metric that tests whether generated answers accurately capture all relevant obligations while avoiding contradictions.

1 Introduction

Regulatory documents are formal semi-legal texts issued by governmental regulatory bodies, setting out rules, guidelines, and standards which organizations must follow to ensure compliance. These documents typically cover a wide range of topics, from environmental and financial compliance to workplace safety and data protection. Characterized by their complex but precise language, they are designed to cover all potential scenarios, but can be challenging to interpret without specialized knowledge. Additionally, regulatory documents are frequently updated to reflect new laws, technological changes, or shifts in societal norms, necessitating continuous monitoring and adaptation by organizations to meet new compliance requirements. Compliance with these regulations requires considerable investment in terms of time and expertise. For example, between 1980 and 2020, the U.S. public and private sectors spent an estimated 292.1 billion hours complying with 36,702 regulations, accounting for approximately 3.2% of total annual working hours (Kalmenovitz, 2023). Moreover, compliance errors can lead to severe penalties, as, for example, demonstrated by the recent C$7.5 million fine levied against the Royal Bank of Canada by FINTRAC for failing to report suspicious transactions (Bloomberg, 2023). Regulatory Natural Language Processing (RegNLP) is a recent multidisciplinary field designed to simplify access to and interpretation of regulatory rules and obligations, reducing errors and enhancing operational efficiency.

Question-answering systems for regulatory documents would enable stakeholders to better navigate complex rules and obligations. They need to extract relevant obligations from these documents to address queries like, *"What steps must we take when a transaction is suspected of involving money laundering?"*. However, due to the intricate structure of regulatory texts and the importance of compliance, generating accurate and comprehensive answers requires successful retrieval and synthesis of obligations. Answers must include all relevant obligations avoiding omissions or contradictions (see Appendix A.1 for more detailed discussion of the steps involved in processing this example query).

This paper makes four significant contributions to RegNLP:

1. **Automated Question-Passages Generation**: We introduce a framework that employs a Large Language Model (LLM) to automatically generate questions, with Natural Language Inference (NLI) integrated at the valida-

tion stage. This framework can be applied to specialized regulatory document collections to create datasets for the development and evaluation of question-answering and single passage or multi passages retrieval systems.

2. **ObliQA - Obligatory Question-Answer dataset**: We develop a multi-document, multi-passage QA dataset[1] specifically to facilitate research in RegNLP. It consists of 27,869 questions and their associated source passages, all derived from the full collection of regulatory documentation provided by Abu Dhabi Global Markets (ADGM)[2], the authority overseeing financial services in the Emirate of Abu Dhabi international financial centre and free zone.

3. **RIRAG - Regulatory Information Retrieval and Answer Generation**: We introduce a task that first involves passage retrieval given a specific regulatory query, identifying all relevant obligations distributed across all regulatory documents. Following this, the task requires generating comprehensive answers based on the retrieved passages. These answers must synthesize information from multiple sources, ensuring that all pertinent obligations are included to fully address the regulatory query. The goal is to create responses that are accurate, complete and concise, providing clear guidance without overwhelming the user with extraneous information.

4. **RePASs - Regulatory Passage Answer Stability Score**: We propose RePASs[3], a novel reference-free metric designed to evaluate the quality of generated answers in regulatory compliance contexts. RePASs is tailored to the specific RegNLP QA task, testing that all relevant obligations are accurately captured, fully reflected in the answers, and devoid of contradictions. To evaluate whether all obligations are covered, we developed an obligation classifier[4] model fine-tuned on LegalBERT (Chalkidis et al., 2020), optimized for the detection of obligatory sentences within regulatory passages.

[1] https://github.com/RegNLP/ObliQADataset
[2] https://www.adgm.com/
[3] https://github.com/RegNLP/RePASs
[4] https://github.com/RegNLP/ObligationClassifier

This paper is structured as follows. Section 2 discusses existing work on RegNLP. Section 3 describes the creation of the ObliQA dataset, and Section 4 introduces the RePASs evaluation metric. Section 5 describes our 2-step baseline Regulatory Information Retrieval and Answer Generation (RIRAG) system, which first retrieves relevant passages for a given query and secondly generates an answer from these passages, and presents our evaluation of both steps. Section 6 concludes the paper and provides directions for future work. Section 7 addresses the limitations of this work.

2 Related Work

2.1 Research on Regulatory NLP

Recent RegNLP research has advanced significantly using various methodologies and datasets. Lau et al. (2005) utilized information retrieval and extraction on accessibility regulations from the US and Europe, formatted into unified XML for semi-structured data handling. Kiyavitskaya et al. (2008) employed the Cerno framework for semantic annotation to automate extracting rights and obligations from U.S. HIPAA and Italian laws. Chalkidis et al. (2018) introduced a hierarchical BiLSTM model for extracting obligations from legal contracts, demonstrating superior performance over traditional models. Nair et al. (2018) developed a pipeline for annotating global import-export regulations, enhancing compliance workflows. Chalkidis et al. (2021) proposed a two-step document retrieval process for EU/UK legislative compliance using BERT-based models, handling long document queries. Abualhaija et al. (2022) presented an automated question-answering system using BERT, achieving high accuracy in identifying relevant texts from European regulations, including the General Data Protection Regulation.

Despite these advances, regulatory document retrieval faces challenges such as implicit obligations, multi-document queries, and cross-referenced clauses. Current systems often struggle to synthesize comprehensive answers when obligations are dispersed or fragmented, and they inadequately address contradictions and coverage. Our work addresses these gaps by introducing a novel dataset and QA task tailored to the financial domain, built from the comprehensive collection of ADGM regulatory documents. Unlike previous datasets that focus on narrower aspects of compliance, our dataset captures a broader range of obligations across mul-

tiple documents. We also propose the RePASs evaluation metric, which goes beyond relevance to assess obligation coverage, confabulation and contradiction, areas where existing metrics fall short.

2.2 Synthetic Data Generation with LLMs

Recent research studies have demonstrated the versatile capabilities of LLMs in enriching QA datasets across diverse domains. Alberti et al. (2019) pioneered methods for generating synthetic QA corpora, using roundtrip consistency for validation and achieving state-of-the-art performance on tasks such as SQuAD2 (Rajpurkar et al., 2018) and NQ (Kwiatkowski et al., 2019). Maatouk et al. (2023) introduced a zero-shot learning approach for neural passage retrieval, using synthetic question generation and hybrid term-neural models to enhance retrieval performance without extensive domain-specific data. Abdullin et al. (2023) focused on synthetic dialogue dataset generation, automating dialogues for linear programming problems using LLMs, thereby aiding in training conversational agents. Collectively, these studies underscore the transformative potential of LLMs for enhancing QA datasets tailored to specific domains.

2.3 Retrieval-Augmented Generation

RegNLP can benefit from advances in retrieval-augmented generation (RAG) (Lewis et al., 2020) and other related technologies to simplify access to regulatory documents and improve compliance. Regulatory documents present two key challenges for application systems: (1) frequent updates that can quickly render systems outdated, and (2) the complexity of specialized terminology and cross-document references that demand a comprehensive understanding of the entire document collection. RAG has significantly improved the accuracy, efficiency, and trustworthiness of LLMs by integrating external, contextually relevant and up-to-date information. Notable approaches include: Self-RAG (Asai et al., 2023) enhances response quality by incorporating self-reflection mechanisms; RAFT (Zhang et al., 2024) optimizes domain-specific knowledge retrieval by filtering out irrelevant data and improving reasoning chains; RobustRAG (Xiang et al., 2024) and RQ-RAG (Chan et al., 2024) tackle challenges such as noisy retrieval results and ambiguous queries; PipeRAG (Jiang et al., 2024a) improves performance by reducing retrieval latency. However, these studies do not consider regulatory documents, so we are interested in testing the ability of RAG methods for solving the RegNLP QA task.

3 ObliQA: Obligation Based Question Answering Dataset

The ObliQA - The Obligation Based Question Answering Dataset creation process involves three main steps: Data Collection, Question Generation, and Question-Passages Validation using natural language inference (NLI), as illustrated in Figure 1.

3.1 Data Collection

The regulatory document collection provided by ADGM consist of 40 documents totaling 640k words, each ranging from 30 to 100 pages. These documents are carefully structured with numbered clauses and subsections to maintain a hierarchical format. More details on document standardization and structuring processes can be found in the Appendix A.2.

3.1.1 Question Generation

Regulatory documents contain extensive descriptions and titles, but for question generation, only passages containing obligations are considered, ensuring the questions are grounded in core compliance requirements.

Input Definition:

Inputs for Single Passage Questions: Inputs for single passage questions isolate one specific source passage per question, simplifying the context and ensuring the relevance of the generated question.

Inputs for Multi-Passage Questions: For broader queries, inputs encompass multiple source passages. Sixty-nine distinct topics related to finance and regulatory compliance (see Appendix A.3 for a sample) guide the aggregation of related regulations into meaningful groups. These groups then form the basis for multi-passage questions by randomly sub-grouping a specified number of rules per question.

GPT-4 Generator: The *gpt-4-turbo-1106* model is employed to generate both questions and answers.

For single passage questions, the prompt is:

> "Your task is to generate realistic and applied questions that pertain to the provided regulatory or compliance material. Ensure that the context implicitly contains the answer to the question."

For multi-passage questions, the complexity increases, and the prompt is adjusted accordingly:

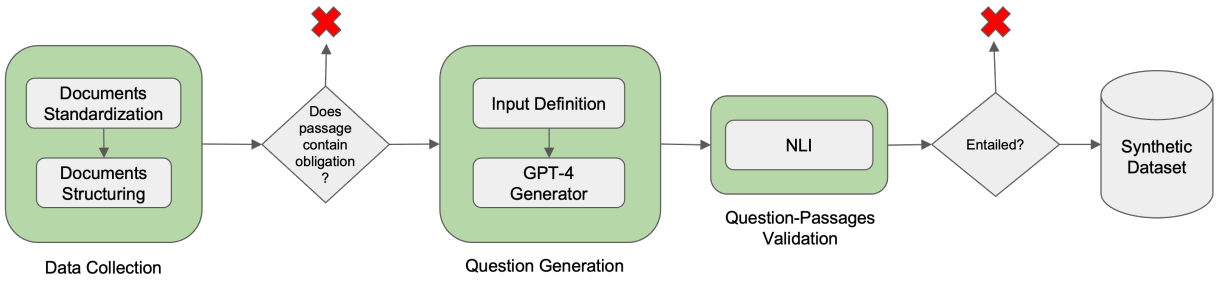

Figure 1: AI-driven framework for generating question-passages pairs to create ObliQA: Obligation-Based Question Answering dataSet for regulatory compliance. This flowchart presents the process from data collection and standardization to question generation and NLI-based validation for constructing a synthetic dataset.

	#Question	1 Passage	2 Passages	3 Passages	4 Passages	5 Passages	6 Passages
Group 1	7988	7988					
Group 2	7254	5432	1822				
Group 3	5365	3385	1434	546			
Group 4	4558	3057	1102	305	94		
Group 10	2704	1325	678	345	174	121	61
Total	**27869**	**21187**	**5036**	**1196**	**268**	**121**	**61**

Table 1: Distribution of questions by number of passages in the ObliQA dataset. Groups indicate the number of input passages.

> "Imagine you are preparing a list of questions that a compliance officer at a company would ask regulatory authorities like the ADGM to clarify their understanding of compliance requirements and to ensure adherence to regulatory standards. Each question should be direct, practical, and structured to obtain detailed information on specific regulatory rules. Avoid formulations that imply the company is describing its own processes, and instead frame questions that seek clarity on regulatory expectations."

3.2 Question-Passage(s) Validation via NLI

Validating the generated question-passage(s) pairs using Natural Language Inference (NLI) involves assessing the semantic relationship between each question and its corresponding source passage(s) to ensure alignment (Dagan et al., 2006). Passage-based validation is employed because obligations are often distributed across multiple sentences, requiring a holistic approach to capture the full context and ensure accurate alignment with the question. The validation process uses the *nli-deberta-v3-xsmall* model with passage(s) as the premise and question as the hypothesis. For each passage associated with a question, NLI is applied to evaluate the relationship. The validation results are categorized as follows.

- **Entailment**: The passage entail the questions being retained.
- **Contradictions**: The passage that contradict the question is eliminated.
- **Neutral**: The neutral passage is retained if the absolute difference between its neutral score and the entailment score is smaller than the difference between its neutral score and the contradiction score; otherwise, it is discarded.

The final dataset consists of approximately 21k questions linked to single passages and 5k questions linked to multiple passages. The distribution of questions among different groups, indicating the number of input passages, is detailed Table 1. The further distribution between the training, testing and development sets is shown in Appendix A.4. A sample of the dataset is provided in Appendix A.5.

3.3 Evaluation of ObliQA

ADGM regulatory experts agreed to evaluate 991 questions and their associated passages from the ObliQA dataset for the purposes of this research, as shown in Table 2. All questions were labeled as an "Expected Question," confirming alignment with real-world compliance scenarios. The evaluation covered 1,847 passages, of which 1,605 (86.91%) were marked as relevant, 65 (3.52%) as irrelevant, and 8 (0.43%) as null. Additionally, 169 passages (9.14%) were identified as relevant but did not directly answer the corresponding questions. These results support the generally high quality of the ObliQA dataset for research in regulatory compliance.

	#Questions	#Passages	Relevant	Irrelevant	Null	Relevant but Indirect
Group 1	79	79	78	1	0	0
Group 2	178	280	194	37	2	47
Group 3	230	428	356	14	6	52
Group 4	245	490	471	7	0	12
Group 10	259	570	506	6	0	58
Total	**991**	**1847**	**1605**	**65**	**8**	**169**

Table 2: Expert evaluation of passage relevance in the ObliQA dataset, categorized by groups.

4 RePASs: Regulatory Passage Answer Stability Score

We propose a reference-free evaluation metric, RePASs, to evaluate generated answers in regulatory compliance contexts. Inspired by recent work on answer attribution to specific sources (Bohnet et al., 2023; Yue et al., 2023), this metric evaluates answers based on three key aspects:

- Every answer sentence must be supported by at least one sentence in the source passage(s).
- The answer must not contain any sentences that contradict the information in the source passage(s).
- The answer must cover all the obligations present in the source passages, meaning that all regulatory obligations should be reflected in the answer.

This metric prefers answers that are not only accurate but also comprehensive in covering all relevant regulatory obligations. The metric includes the following steps:

Step 1- Generation of Entailment and Contradiction NLI Pair Matrices: We generate an NLI pair matrix by comparing each passage sentence (premise) with each answer sentence (hypothesis) using the *cross-encoder/nli-deberta-v3-xsmall* (He et al., 2021a). The model provides three outputs for each pair: probabilities for entailment, contradiction, and neutrality. These probabilities are organized into two matrices: the entailment matrix, which contains the entailment probabilities for each passage-answer sentence pair, and the contradiction matrix, which stores the contradiction probabilities for each sentence pair.

Step 2- Calculation of Entailment and Contradiction Scores: We reduce the size of the entailment and contradiction matrices by selecting the highest probability for each answer sentence across the passage sentences. Matrix reduction ensures that only the most relevant support from or conflict with the source passage is considered. The entailment score is computed by averaging the highest entailment probabilities across all answer sentences. Similarly, the contradiction score is obtained by averaging the highest contradiction probabilities.

$$E_s = \frac{1}{N} \sum_{i=1}^{N} \max_j P_{\text{entailment}}(p_j, a_i) \quad (1)$$

$$C_s = \frac{1}{N} \sum_{i=1}^{N} \max_j P_{\text{contradiction}}(p_j, a_i) \quad (2)$$

Where E_s is the Entailment Score, and C_s is the Contradiction Score. N represents the total number of sentences in the generated answer. $P_{\text{entailment}}(p_j, a_i)$ is the probability that the i-th sentence in the answer (a_i) is entailed by the j-th sentence in the source passage (p_j). The $\max$ function selects the highest entailment probability for each answer sentence (a_i) across all sentences in the source passage (p_j). Similarly, $\max_j$ selects the highest probability for each answer sentence across all sentences in the source passage.

Step 3- Calculation of Obligation Coverage Score: The obligation coverage score evaluates how accurately the generated answer reflects the obligations present in the source passage(s). First, a dataset is created using an automated pipeline powered by the *GPT-4-turbo-1106* model, which extracts and labels sentences from regulatory documents as either obligations or non-obligations using zero-shot learning. This labeled dataset is then used to fine-tune a LegalBERT-based classifier model (Chalkidis et al., 2020), which is optimized for detecting obligatory sentences in regulatory texts. Both the source passages and generated answers are tokenized into sentences to enable a detailed, sentence-level comparison for obligation coverage.

For each obligation sentence detected in the source passage, the system verifies its coverage in the generated answer. This is done by comparing the obligation sentences with those in the answer

using the NLI model *microsoft/deberta-large-mnli* (He et al., 2021b). If any sentence in the answer demonstrates an entailment score exceeding 0.7 when compared to an obligation sentence from the source passage(s), the obligation is considered covered. The obligation coverage score is calculated as the ratio of the number of covered obligations to the total number of obligations present in the source passage(s).

The formula for the Obligation Coverage Score is given by:

$$OC_s = \frac{1}{M} \sum_{k=1}^{M} \mathbb{1}\left(\max_{l} P_{\text{entailment}}(o_k, a_l) > 0.7\right) \quad (3)$$

Where OC_s represents the Obligation Coverage Score, and M is the total number of obligation sentences in the source passage. $P_{\text{entailment}}(o_k, a_l)$ denotes the probability that the k-th obligation sentence in the source passage (o_k) is entailed by the l-th sentence in the answer (a_l). The $\max_l$ function selects the highest entailment probability for each obligation sentence across all sentences in the answer. The indicator function $\mathbb{1}$ returns 1 if the highest entailment score exceeds 0.7, meaning the obligation is considered covered.

Step 4- Calculation of the Regulatory Passage Answer Stability Score: The final step integrates the three components into a single composite score. The score penalizes contradictions and normalizes the result between 0 and 1.

$$RePASs = \frac{\mathrm{E}_s - \mathrm{C}_s + \mathrm{OC}_s + 1}{3} \quad (4)$$

Traditional metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), FactCC (Kryscinski et al., 2020), and SummaC (Laban et al., 2022) are used to evaluate text generation but fall short in RIRAG contexts. BLEU and ROUGE focus on n-gram overlap and lexical similarity, neglecting factual accuracy and the completeness of regulatory obligations. BERTScore captures semantic similarity but lacks tools to assess whether answers cover all obligations or avoid contradictions. While FactCC ensures factual consistency, it doesn't verify that all necessary regulatory obligations are included. SummaC evaluates consistency with the source text but doesn't account for regulatory coverage or contradictions. In contrast, RePASs is tailored specifically to regulatory compliance. It evaluates answers based on entailment, contradiction, and coverage, ensuring that responses not only maintain coherence but also comply with the task requirements.

Metric	Sub-metric	Score
Standard Evaluation Metrics		
SummaC	-	0.5890
FactCC	-	0.8571
BERTScore	Precision	0.6710
	Recall	0.7449
	F1	0.7080
BLEU	-	0.8300
ROUGE-1	Precision	0.8606
	Recall	1.0000
	F1	0.9218
ROUGE-2	Precision	0.8518
	Recall	1.0000
	F1	0.9155
ROUGE-L	Precision	0.8606
	Recall	1.0000
	F1	0.9218
RePASs	E_s	0.7639
	C_s	0.3336
	OC_s	1.0000
	RePASs	0.8101

Table 3: Performance of evaluation metrics on the gold standard dataset

To assess RePASs, we tested it on real-world questions using expert-provided obligations and answers from ADGM. Table 3 shows that while ROUGE-L achieves a high F1 score (0.9218), it doesn't ensure obligation coverage or the absence of contradictions. Similarly, BERTScore's F1 score (0.7080) reflects semantic similarity but fails in regulatory compliance. FactCC's score (0.8571) emphasizes factual consistency but doesn't capture regulatory requirements. SummaC (0.5890) also falls short of the precision needed in these contexts. Table 3 illustrates that RePASs aligns well with expert judgments, and the results provide an indicative upper bound for the performance of an automated system.

5 Regulatory Information Retrieval and Answer Generation Task

The Regulatory Information Retrieval and Answer Generation (RIRAG) task includes two key subtasks: passage retrieval and answer generation. In the passage retrieval subtask, the system retrieves all relevant obligations from regulatory documents in response to a given question. The goal is to ensure that all pertinent obligations are identified and collected. In the answer generation subtask, the system synthesizes the retrieved information into a precise, informative response that comprehensively

addresses the question. In this section, we describe a baseline approach for RIRAG, setting a standard that will support future research in the field.

5.1 Passage Retrieval

We experiment with six retrieval models, which are (1) **BM25** (Robertson et al., 1994): the traditional lexical-based model; (2) **DRAGON+** (Lin et al., 2023): the State-of-the-Art (SotA) single-vector dense retriever model fine-tuned on MS MARCO (Nguyen et al., 2016); (3) **SPLADEv2** (Formal et al., 2021): the SotA neural sparse retriever fine-tuned on MS MARCO; (4) **ColBERTv2**: the SotA multi-vector dense retriever model finetuned on MS MARCO; (5) **NV-Embed-v2** (Lee et al., 2024): the SotA single-vector dense retriever model fine-tuned on multiple text-embedding datasets, achieving the best[5] performance on Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023); (6) **BGE-EN-ICL** (Xiao et al., 2023): the SotA single-vector dense retrieval model fine-tuned on multiple text-embedding datasets, achieving the second-best[6] performance on MTEB.

DRAGON+, SPLADEv2 and ColBERTv2 represent the three main architectures for neural retrieval, achieving SotA effectiveness with limited training resources and parameters (all around 110M parameters). NV-Embed-v2 and BGE-EN-ICL are the best-performing dense retrievers when the training resources are extensive and the parameters are largely scaled up (7.1B and 7.9B parameters, respectively). Additionally, BGE-EN-ICL supports In-Context Learning (ICL) (Brown et al., 2020) for text embeddings, providing an option for obtaining task-adapted model with a low cost via adding example demonstrations in the model input.

Since the retrieval task in RIRAG searches for passages from a corpus of long documents, we also consider it as a Document-Aware Passage Retrieval (DAPR) (Wang et al., 2024) task to jointly model the document context and passage retrieval. To this end, we apply **rank fusion** to linearly fuse the passage ranking by the neural or BM25 retrievers and the document ranking by the BM25 retriever. Wang et al. (2024) shows that this approach yields a consistent improvement on retrieval tasks across different domains. The details are available in Appendix A.7. The counterpart without rank fusion is denoted as **passage-only**.

[5]As of Aug 30, 2024.

[6]As of Aug 30, 2024.

For evaluating the retrieval module in RIRAG, we use recall@10 as the main evaluation metric, as we rely on the retrieval module to cover the relevant information as much as possible while leaving the burden of noise filtering to the answer-generation module. This approach is in line with LongRAG (Jiang et al., 2024b). To additionally assess the ranking ability within the top-10 passages, MAP@10 (Mean Averaged Precision) is also used for diagnostic purposes.

For the experimental setup, we truncate the queries and the passages to 512 tokens. In rank fusion, we fuse the ranking of top-100 passages and the ranking of all the documents. The fusion weight on the document side is set to 0.1. The ranking scores are normalized (cf. Equation 5) with respect to the top-100 passages into the range of $[0, 1]$ for filtering the passages conveniently in the answer-generation step. For BGE-EN-ICL, we consider the scenarios of zero-shot, 1-shot, 3-shot and 5-shot. The ICL examples are obtained via random sampling from the training set of ObliQA. According to their recommended usage, the instructions/query prefixes[7] in NV-Embed-v2 and BGE-EN-ICLL are "Given a question, retrieve passages that answer the question" and "Given a financial-regulation query, retrieve relevant passages that answer the query", respectively.

5.2 Answer Generation

The answer generation process begins once 10 relevant passages have been retrieved. In our post-retrieval stage, we implement a score-based filtering approach with a threshold of 0.2 to identify significant drops in relevance between consecutive passages. We also enforce a minimum score criterion of 0.7, ensuring that only the most relevant passages are considered for generating answers. Based on these filtered passages, we use the *gpt-4-turbo-1106* model to generate comprehensive answers. The model follows a tailored prompt designed to simulate the role of a regulatory compliance assistant, integrating all relevant obligations and best practices from the passages into a cohesive response. The prompt reads:

> "You are a regulatory compliance assistant. Provide a detailed answer for the question that fully integrates all the obligations and best practices from the given passages. Ensure your response is cohesive and directly addresses the question. Synthesize the information from all passages into a single, unified answer."

[7]These models do not support instructions for the passage input.

For evaluation, the responses are organized in JSON format, as shown in the Appendix A.6.

5.3 Results

5.3.1 Passage Retrieval Results

The passage retrieval results on the test split of ObliQA are shown in Table 4. We find the models of DRAGON+, SPLADEv2 and ColBERTv2, which are fine-tuned with limited training data (i.e. MS MARCO only) and limited parameters, generalize poorly, achieving worse or on-par recall@10 and MAP@10 than simple BM25. ColBERTv2, the best model in this group, outperforms BM25 by 0.6 points MAP@10 and 1.1 points recall@10. On the other hand, the models of NV-Embed-v2 and BGE-EN-ICL, which are fine-tuned with more training data and more parameters, outperform BM25 significantly in all the settings. Compared with BM25, recall@10 and MAP@10 are improved by up to 3.4 points and up to 2.1 points, respectively. These results show that scaling the retrievers with more training data and parameters is necessary for good generalization on the retrieval task in RIRAG. We find ICL is important for adapting the neural models to this new task, while the advantage of including more examples in the demonstration is very marginal. Lastly, we find that rank fusion, which considers the document context during passage retrieval, improves the performance only marginally. We assume this is because the query generation process only considers the target passages as input, resulting in fairly self-contained query-evidence pairs.

Model	Passage-only		Rank fusion	
	R@10	M@10	R@10	M@10
BM25	76.1	62.4	76.4	62.5
DRAGON+	74.3	58.1	74.5	58.2
SPLADEv2	75.7	60.4	75.9	60.7
ColBERTv2	77.7	63.0	77.7	63.2
NV-embed-v2	78.3	63.5	78.2	63.6
BGE-EN-ICL (0-shot)	77.1	62.4	77.4	62.5
BGE-EN-ICL (1-shot)	79.2	64.1	79.3	64.2
BGE-EN-ICL (3-shot)	**79.5**	64.1	**79.5**	64.2
BGE-EN-ICL (5-shot)	79.3	**64.4**	**79.5**	**64.5**

Table 4: Results of the retrieval task on the test dataset. R@10 and M@10 represent Recall@10 and MAP@10, respectively.

5.3.2 Answer Generation Results

Table 5 illustrates the performance differences between BM25(passage-only)+GPT-4 and BM25(rank fusion)+GPT-4 used in our answer generation task. Similar to the retrieval task results, where rank fusion showed slight performance improvements, we observe that BM25 (rank fusion) + GPT-4 outperforms BM25 (passage-only) + GPT-4 across all metrics. Specifically, rank fusion achieves higher E_s (0.320 vs. 0.308), C_s (0.131 vs. 0.123), OC_s(0.222 vs. 0.214), and RePASs (0.470 vs. 0.466). This highlights the connection between better passage retrieval and improved answer generation in regulatory tasks.

Method	E_s	C_s	OC_s	RePASs
BM25(passage-only)+GPT-4	0.308	0.123	0.214	0.466
BM25(rank fusion)+GPT-4	**0.320**	**0.131**	**0.222**	**0.470**

Table 5: Results of the answer generation task using RePASs on the ObliQA test dataset. E_s, C_s, OC_s and RePASs represent Entailment, Contradiction, Obligation Coverage and RePAS score, respectively.

6 Conclusion

In conclusion, this paper advances the field of RegNLP by introducing research infrastructure and metrics aimed at improving the overall efficiency and accuracy of regulatory compliance. The automated question-passage generation framework and the ObliQA dataset contribute to the study of the retrieval and understanding of regulatory information. Furthermore, the regulatory information retrieval and answer generation task, supported by the novel RePASs evaluation metric, provides a framework for improving on our baseline approach.

In the future, the field of RegNLP could expand with the integration of new tasks. For instance, *Summarization and Simplification* methods could be used to develop tools to make complex regulatory texts simpler and more accessible to non-experts. *Automated Compliance Checking* could leverage cross-document analysis techniques to improve regulatory adherence by comparing an organization's internal compliance documents with regulatory documents. *Regulatory Gap Analysis* could focus on identifying vagueness, ambiguity or contradictions in regulations to aid in their refinement and ongoing compliance when regulations are added or updated. The continued development of these tasks will equip RegNLP to better serve the needs of regulatory bodies and regulated organizations alike, making compliance a more streamlined, reliable and efficient process.

7 Limitations

One significant limitation we encountered is the scarcity of publicly available datasets. Regulatory compliance often involves handling highly sensitive and confidential information, such as personal data or financial records, which presents challenges in terms of data privacy and security. To address this, we created a semi-synthetic dataset—ObliQA—using our Automated Question-Passage Generation Framework. During the validation phase, we faced difficulties stemming from multidisciplinary complexity and shortage of experts. However, experts evaluated a small subset of the dataset and confirmed that 100% of the generated questions met expectations, while 86% of the answer passages were deemed relevant to the questions.

Another key limitation is the lack of standardized formats for regulatory documents. Regulatory documents vary significantly across industries, regions, and even within the same organization, making it difficult to develop models capable of generalization. In some cases, even within the same regulatory authority, documents are prepared in different formats due to departmental variations. In this work, standardizing 40 documents required over 300 person-hours of effort. Additionally, the handling of tables and figures presented further challenges; tables were semi-automatically converted to text format, and figures were interpreted using the GPT-4 model.

Furthermore, the absence of comprehensive benchmarks is another limitation. There is no standardized benchmarking system available to evaluate the performance of RegNLP models across various regulatory domains, making it challenging to assess the effectiveness of different approaches or track progress in the field. Although we proposed the RePASs metric, we were unable to compare it with other models due to the lack of obligation coverage in existing benchmarks. Nevertheless, on genuine gold standard data from ADGM, RePASs demonstrated its suitability for RIRAG tasks.

Acknowledgement

We would like to express our gratitude to the Financial Services Regulatory Authority (FSRA)'s Applied Innovation & Research Team within ADGM for providing data and annotations. We also gratefully acknowledge ADGM's support through a grant, enabling access to OpenAI GPT models via the Azure cloud, which contributed significantly to the technical development of this research.

References

Yelaman Abdullin, Diego Molla, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2023. Synthetic dialogue dataset generation using LLM agents. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 181–191, Singapore. Association for Computational Linguistics.

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.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

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

Bloomberg. 2023. Rbc fined $5.5 million for breaking canadian money laundering rules. Accessed: 2023-05-20.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and modeling for attributed large language models. *Preprint*, arXiv:2212.08037.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2018. Obligation and prohibition extraction using hierarchical RNNs. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 254–259, Melbourne, Australia. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

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. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3498–3511, Online. Association for Computational Linguistics.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *Preprint*, arXiv:2404.00610.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse lexical and expansion model for information retrieval. *ArXiv preprint*, abs/2109.10086.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

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

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

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024b. Longrag: Enhancing retrieval-augmented generation with long-context llms. *Preprint*, arXiv:2406.15319.

Joseph Kalmenovitz. 2023. Regulatory Intensity and Firm-Specific Exposure. *The Review of Financial Studies*, 36(8):3311–3347.

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*, pages 154–168, Berlin, Heidelberg. Springer Berlin Heidelberg.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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*, ICAIL '05, page 146–154, New York, NY, USA. Association for Computing Machinery.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

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. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400, Singapore. Association for Computational Linguistics.

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

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

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, Santa Fe, New Mexico. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 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).

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2024. DAPR: A benchmark on document-aware passage retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4313–4330, Bangkok, Thailand. Association for Computational Linguistics.

Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. Bert-based dense retrievers require interpolation with BM25 for effective passage retrieval. In *ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021*, pages 317–324. ACM.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust rag against retrieval corruption. *Preprint*, arXiv:2405.15556.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. Raft: Adapting language model to domain specific rag. *Preprint*, arXiv:2403.10131.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Appendix

A.1 Illustrative Example of RIRAG for Compliance Queries

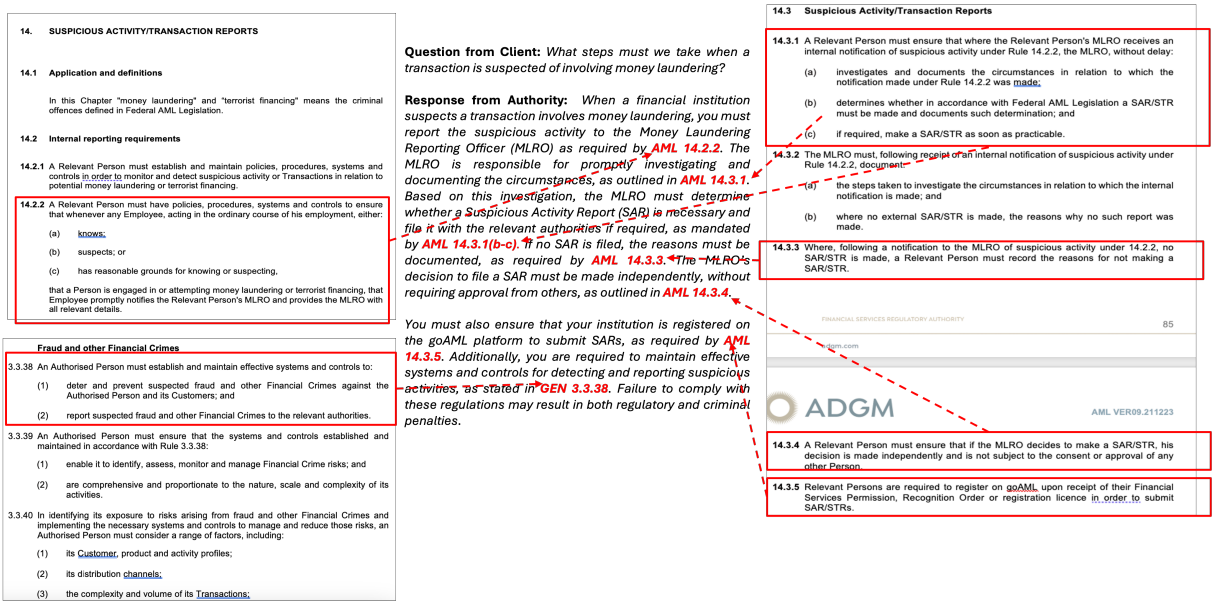

Figure 2: An example of a comprehensive response by retrieving and synthesizing relevant passages from regulatory documents in response to a client question

Figure 2 illustrates a RIRAG task. In response to a client's query about handling suspicious financial transactions, the system retrieves key regulations from several passages across the Anti-Money Laundering (AML) and General (GEN) sections of regulatory documents. By synthesizing these passages, the system generates a concise, comprehensive answer, helping streamline the institution's compliance with complex regulations.

A.2 Document Standardization and Structuring

Documents Standardization: To ensure the integrity of the dataset, the documents underwent a meticulous semi-automatic restructuring process. This step is critical due to the diverse formats employed by different departments in the preparation of regulations. Each document is parsed and transformed into .txt format. Tables within the documents are delineated with *Table Start* and *Table End* tags, while graphical content is transcribed into text using manually-corrected GPT-4 output and encapsulated with *Figure Start* and *Figure End* tags.

Structuring the Data: Subsequently, the standardized .txt files are converted into a structured JSON format for data manipulation and analysis. Below is an example of the JSON structure:

```
{
    "ID": "2230472f-a9d4-4b81-842f-964c0653f8e7",
    "DocumentID": 20,
    "PassageID": "1.1",
    "Passage": "This Guidance has been produced to help legal practitioners in ADGM to understand the intention behind the
        ↪ Application of English Law Regulations 2015 (the 'Application Regulations') and how English law has been implemented
        ↪ in ADGM..."
}
```

The `PassageID` is used to represent the hierarchical information of the passages within the documents.

A.3 Keyword-Topic List for ObliQA Dataset

```
topics_keywords = {
    "AML Compliance": [
        "money laundering", "compliance program", "regulatory requirements"
    ],
    "Anti-Money Laundering": [
        "AML","money laundering", "KYC", "know your customer", "financial crime", "terrorist financing",
        "due diligence", "suspicious activity", "compliance program"
```

```
    ],
    "Audit and Monitoring": [
        "audit", "compliance monitoring", "internal audits", "external audits"
    ],
    "Blockchain Technology": [
        "blockchain", "blockchain technology", "smart contract", "tokenization"
    ],
    "Blockchain-based Securities": [
        "blockchain-based securities", "blockchain technology in securities"
    ],
    ....
    "Virtual Asset Regulation": [
        "virtual assets", "crypto assets", "digital asset regulation", "virtual asset service providers", "VASP", "crypto
            ↪ exchanges", "crypto custodians", "ICO regulations", "token classifications"
    ]
}
```


A.4 Distribution of questions in the ObliQA dataset

	#Question	1 Passage	2 Passages	3 Passages	4 Passages	5 Passages	6 Passages
train	**22295**	16946	4016	975	202	100	56
test	**2786**	2126	506	105	36	9	4
development	**2888**	2215	514	116	30	12	1

Table 6: Distribution of questions in the ObliQA dataset across training, testing, and development sets , categorized by the number of associated passages.

A.5 JSON Data Examples from the ObliQA Dataset


```
[
  {
    "QuestionID": "739921c1-385a-4735-a052-dee9fba73602",
    "Question": "What are the key compliance indicators that a Fund Manager should monitor to confirm that a Passported Fund is
        ↪ being managed and operated within its constitutional framework and applicable ADGM legislation?",
    "Passages": [
      {
        "DocumentID": 16,
        "PassageID": "Part 3.6.(2)",
        "Passage": "Each Reporting UAE Financial Institution shall establish and implement appropriate systems and internal
            ↪ procedures to enable its compliance with the Cabinet Resolution and these Regulations."
      },
      {
        "DocumentID": 5,
        "PassageID": "6.1.2",
        "Passage": "The Fund Manager of a Passported Fund must:\n(a)\tensure that the Passported Fund is at all times managed and
            ↪ operated in compliance with its constitution, in accordance with applicable ADGM legislation, and with these
            ↪ Rules; and\n(b)\tmaintain, or cause to be maintained, a Unitholder register for the Passported Fund."
      }
    ]
  },
  {
    "QuestionID": "8dc451a4-ec55-4fa5-abce-b0b8764a9338",
    "Question": "Can the Regulator provide case studies or examples where the conduct of an Approved Person has been deemed
        ↪ compliant or non-compliant with the Principles, to help clarify the practical application of these rules?",
    "Passages": [
      {
        "DocumentID": 6,
        "PassageID": "PART 5.13.3.9.Guidance.3",
        "Passage": "The onus will be on the Regulator to show that he is culpable, taking into account the standard of conduct
            ↪ required under the principle in question. In determining whether or not the particular conduct of an individual
            ↪ complies with the principles, the Regulator will take into account whether that conduct is consistent with the
            ↪ requirements and standards relevant to an individual's role and the information available to him."
      },
      {
        "DocumentID": 7,
        "PassageID": "2.3.1.Guidance.3.",
        "Passage": "In those circumstances, the onus is on the Regulator to show that the Approved Person is culpable, taking into
            ↪  account the standard of conduct required under the Principle in question. In determining whether or not the
            ↪ particular conduct of an Approved Person complies with the Principles for Approved Persons, the Regulator will
            ↪ take account of whether that conduct is consistent with the requirements and standards relevant to their
            ↪ Authorised Person, their own role and the information available to them.\n"
      }
    ]
  }
]
```

A.6 JSON Data Example for Evaluation of Answer Generation Step


```
{
    "QuestionID": "12345678-abcd-1234-efgh-ijklmnopqrst",
    "Question": "What are the regulatory requirements for data encryption in financial services?",
    "RetrievedPassages": [
        "Financial institutions must encrypt sensitive data at rest and in transit. Key requirements include the use of AES-256
            ↪ encryption for data at rest and TLS 1.2 or higher for data in transit."
    ],
    "Answer": "Regulatory requirements for data encryption in financial services mandate using AES-256 encryption for data at rest
        ↪  and TLS 1.2 or higher for data in transit to ensure security of sensitive information.",
    "RetrievedIDs": [
        "9abc1234-def5-6789-ghij-klmnopqrstuv"
    ]
}
}
```


A.7 Rank Fusion

Rank fusion fuses the relevance scores from a BM25 retriever and a neural retriever. We compute the fusion as the convex combination of the normalized relevance scores (Wang et al., 2021):

$$s_{\text{convex}}(q, p, d) = \alpha \hat{s}_{\text{BM25}}(q, p) + (1 - \alpha)\hat{s}_{\text{neural}}(q, d),$$

where $\alpha \in [0, 1]$ is the fusion weight, and $\hat{s}_{\text{BM25}}/\hat{s}_{\text{neural}}$ represents the normalized BM25/neural-retrieval relevance score, respectively. The normalization for a relevance score s is calculated as:

$$\hat{s}(q, c) = \frac{s(q, c) - m_q}{M_q - m_q}, \tag{5}$$

where c represents the passage/document candidate, m_q and M_q are the min. and max. relevance scores of the top candidates for q, respectively. For any misaligned candidates, a zero score is taken for the candidate-missing side. In this work, the fusion is applied between the BM25 document rankings and the neural passage rankings of a certain cut-off.