Shared Task RIRAG-2025: Regulatory Information Retrieval and Answer Generation

Tuba Gokhan[1], **Kexin Wang**[2], **Iryna Gurevych**[1,2], **Ted Briscoe**[1]
[1]Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)
[2]Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt

Abstract

This paper provides an overview of the Shared Task RIRAG-2025, which focused on advancing the field of Regulatory Information Retrieval and Answer Generation (RIRAG). The task was designed to evaluate methods for answering regulatory questions using the ObliQA dataset. This paper summarizes the shared task, participants' methods, and the results achieved by various teams.

1 Introduction

Regulatory compliance is a critical but highly complex domain, requiring organizations to interpret and adhere to a wide range of rules, standards, and obligations. These tasks are traditionally labor-intensive and involve meticulous analysis of all regulatory documents to ensure compliance. The growing volume and complexity of regulations has made manual processes increasingly unsustainable. Addressing these challenges necessitates innovative solutions to automate regulatory compliance tasks.

The **Regulatory Information Retrieval and Answer Generation (RIRAG)** focuses on automating two core processes: retrieving relevant regulatory information and generating concise, accurate answers to compliance-related questions. By combining information retrieval and answer generation, RIRAG provides a framework to streamline compliance workflows and enhance organizational efficiency.

To foster collaboration and innovation in this emerging field, we organized the **RIRAG-2025 shared task**. This shared task aims to benchmark and advance methodologies for regulatory information retrieval and answer generation, bringing together academic and industrial researchers to address real-world compliance challenges.

2 RIRAG-2025

2.1 Task Description

The Regulatory Information Retrieval task seeks to automate the extraction and synthesis of information from complex regulatory documents. This involves addressing multi-passage and multi-document challenges inherent to regulatory compliance. The task is divided into two subtasks:

Subtask 1: Information Retrieval: The objective is to retrieve the most relevant passages from a regulatory corpus for a given compliance-related question. These passages form the foundation for generating accurate answers.

Subtask 2: Answer Generation: This subtask focuses on generating a comprehensive based on the passages retrieved in Subtask 1. The generated answers must integrate all relevant obligations while avoiding contradictions or omissions.

2.2 Dataset: ObliQA

The shared task leverages the ObliQA dataset [1], a regulatory compliance-focused dataset derived from Abu Dhabi Global Market (ADGM) regulations. ObliQA comprises 27,869 questions, each annotated with corresponding passages, making it a robust resource for developing and benchmarking RIRAG systems. The dataset poses unique challenges, including:

Single-Passage Questions: Questions that require retrieving and analyzing a single passage.

Multi-Passage Questions: Questions necessitating the integration of multiple passages for a complete answer.

2.3 Baseline System

The baseline system (Gokhan et al., 2024) serves as a foundational framework for the participants, providing a clear reference for addressing the RIRAG

[1] https://github.com/RegNLP/ObliQADataset

Proceedings of the 31st International Conference on Computational Linguistics, pages 1–4
January 19–24, 2025.

task. For passage retrieval, the system combines BM25, dense retrieval models (e.g. DRAGON + and ColBERTv2), and rank fusion techniques to retrieve relevant passages. The answer generation component uses GPT-4-turbo-1106 with prompt engineering to synthesize obligation-focused answers from the retrieved passages.

2.4 Evaluation

To evaluate system performance, different metrics are applied to the two subtasks. For Subtask 1 (Information Retrieval), Recall at 10 (R@10) and Mean Average Precision at 10 (M@10) are used to assess the system's ability to retrieve relevant passages effectively. For Subtask 2 (Answer Generation), the Regulatory Passage Answer Stability Score (RePASs)[2] measures the quality of generated answers based on their entailment with source passages, avoidance of contradictions, and coverage of obligations.

3 Overview of Shared Task

The task was organized in time for COLING 2025 as part of the RegNLP 2025 workshop. A total of 19 teams participated, with 16 of them submitting both their system results and papers describing their approach.

During the development stage, the teams worked with the publicly available ObliQA dataset, which served as the primary resource for system training and fine-tuning. To support additional methodological exploration, the entire set of 40 hierarchically structured regulatory documents, from which the ObliQA dataset was derived, was also made available to participants.

In the evaluation stage, submissions were tested on a hidden subset of the ObliQA dataset consisting of 446 unseen questions. These questions were provided without access to their associated ground truth passages.

4 Overview of Teams' Methodologies

The participating teams in the RIRAG-2025 shared task employed diverse methodologies to address the challenges posed by the two subtasks. This section provides an overview of the approaches used by the teams, categorized by subtask.

[2] https://github.com/RegNLP/RePASs

4.1 Subtask 1: Information Retrieval

The participating teams employed a diverse range of methods for the information retrieval task, combining sparse retrieval, dense retrieval, hybrid systems, and re-ranking strategies to optimize passage retrieval for regulatory queries.

BM25 was a foundational component in many teams' systems, often augmented with additional techniques to enhance performance. Teams utilizing BM25 included USTC-IAT-United, NUST Nova, NUST Alpha, JurisCore, Ocean's Eleven, NLP-MindMappers, NLP-MJR, TEAM: 1-800, Indic aiDias, and AUEB. Hybrid systems were frequently implemented to balance lexical precision with semantic understanding. For example, USTC-IAT-United combined BM25, DRAGON+, ColBERTv2, and a fine-tuned LLaMA-2-7B model, employing a hybrid expert mechanism with dynamic weight assignment. Ocean's Eleven utilized BM25, NV-Embed-v2, and BGE-en-ICL embeddings, leveraging reciprocal rank fusion and NLI-based re-ranking to enhance retrieval relevance. Havelsan integrated bge-m3, e5-large-v2, and Jina embeddings, combined with context-aware chunking, to create a robust hybrid retrieval system.

Many teams further refined retrieval results using re-ranking models and dynamic filtering. For instance, AICOE employed text-embedding-ada-002 embeddings alongside RankGPT for sliding-window re-ranking, while Indic aiDias implemented a multi-stage tuning process with reciprocal rank fusion and context-based filtering. NLP-Alpacas applied msmarco-roberta-base-v2 and BAAI/bge-base-en-v1.5, using triplet-based fine-tuning and FAISS indexing for improved passage ranking.

Table 1 provides an overview of the teams and their respective methods.

4.2 Subtask 2: Answer Generation

The participating teams adopted various methods for the answer generation task, focusing on large language models (LLMs), prompt engineering, and post-processing strategies to produce accurate regulatory-aligned responses. Many teams employed state-of-the-art generative models to synthesize answers from retrieved passages. For instance, NLP-MindMappers and NUST Omega utilized Few-Shot prompting and Chain-of-Thought (CoT) techniques with GPT models to generate structured and comprehensive answers. Mean-

Table 1: Overview of Teams' Methodologies for Subtask 1: Information Retrieval

Paper ID	Team Name	Retrieval Methods	Key Features
11	USTC-IAT-United	LLaMA-2-7B fine-tuned + BM25 + DRAGON+ + ColBERTv2	Hybrid expert mechanism, Dynamic weight assignment
12	NUST Nova	LegalBERT + BM25 + FAISS + Neo4j Graph-Based Retrieval	Graph-based retrieval, Score fusion, Re-ranking
13	NUST Alpha	BM25 + FAISS	Rank fusion, GPT-based filtering , Re-ranking
14	NUST Omega	LegalBERT + Gemini + OpenAI embeddings + FAISS	Metadata-driven query matching, Topic modeling
15	Havelsan	bge-m3 + e5-large-v2 + Jina embeddings + hybrid search	Hybrid retrieval, Context-aware chunking, Re-ranking
16	Obayer	intfloat/multilingual-e5-large + txtai	
17	AICOE	text-embedding-ada-002 + RankGPT	Two-step retrieval, Sliding-window re-ranking
18	JurisCore	BM25 + Dense Retrieval + BDD-FinLegal	Cross-encoder re-ranking, Adaptive dynamic weighting
19	Ocean's Eleven	BM25 + NV-Embed-v2 + BGE-en-ICL	Reciprocal rank fusion, NLI-based re-ranking
20	NLP-MindMappers	BM25 + all-MiniLM-L6-v2 + FAISS	Bi-encoder retrieval, BM25 re-ranking, Multiple negatives ranking loss
21	NLP-Alpacas	msmarco-roberta-base-v2 + BAAI/bge-base-en-v1.5 + FAISS	Multiple negatives ranking loss, Triplet-based fine-tuning, FAISS-based indexing
22	NLP-MJR	BM25 + BAAI/bge-small-en-v1.5	Weighted score fusion, Semantic matching, Hybrid retrieval
23	TEAM: 1-800	BM25 + BGE-small-en-v1.5 + MPNet V2	Lexical-semantic score fusion, LeSeR reranking, MNSR fine-tuning
24	Indic aiDias	BM25 + BGE-EN-ICL + E5-FT + Q2Q	Reciprocal rank fusion, Context-based filtering, Multi-stage tuning
25	AUEB	BM25 + Voyage-Law-2 + Voyage-Finance-2 + Voyage-Rerank-2	Triple rank Fusion, Re-ranking
26	NLP-LingoLlamas	MiniLM-L6-v2 + stella en 400M v5 + Gemini-1.5-pro-002	Fine-tuning with negatives, Inverted re-ranking retrieval

Table 2: Overview of Teams' Methodologies for Subtask 2: Answer Generation

Paper ID	Team Name	Generative Models	Key Features
11	USTC-IAT-United	Qwen2-72B	Scoring-based passage filtering, Prompt
12	NUST Nova	Llama3-70b	Prompt
13	NUST Alpha	GPT-3.5	Prompt
14	NUST Omega	GPT *	Few-Shot, CoT, Prompt
15	Havelsan	LLaMA-3.1-8B-Instruct	Prompt
16	Obayer	—	
17	AICOE	GPT-4o	Prompt
18	JurisCore	—	
19	Ocean's Eleven	LLaMa-3.1-8B-Instruct, CFG, CAD	Prompt
20	NLP-MindMappers	Gemma 2B , GPT-4o	Few-Shot, CoT
21	NLP-Alpacas	T5-base, GPT-4o	Prompt
22	NLP-MJR	GPT 3.5 Turbo, GPT-4o Mini, Llama 3.1	Prompt
23	TEAM: 1-800	Qwen2.5 7B, Gemma-2 9B, Mistral 7B, Nemo 12B	Prompt
24	Indic aiDias	LLaMA-3.1-8B-Instruct, Single line, Identity function	
25	AUEB	GPT-4o Mini	Scoring and Obligation-based passage filtering, Post-Processing
26	NLP-LingoLlamas	Gemini-1.5-pro-002	Prompt

while, AUEB and USTC-IAT-United implemented passage filtering mechanisms to ensure the relevance and alignment of generated responses with regulatory obligations. Table 2 summarizes the models and key features utilized by each team.

5 Teams' Evaluation Results

The evaluation of team submissions was conducted separately for both subtasks.

The evaluation was based on a hidden subset of the ObliQA dataset consisting of 446 unseen questions. Table 3 provides a detailed breakdown of the scores for all teams and their submissions. Some teams submitted multiple versions of their systems, showcasing iterative improvements and different configurations.

Subtask 1 (Information Retrieval): The highest R@10 and M@10 scores were achieved by **Indic aiDias** with their first submission, scoring 0.787 and 0.663, respectively. Teams **NLP-MJR** (R@10: 0.731, M@10: 0.602) and **TEAM: 1-800** (R@10: 0.705, M@10: 0.562) also performed strongly in the retrieval subtask.

Subtask 2 (Answer Generation): The best RePASs score (0.973) was achieved by **Indic aiDias** with their first submission, closely followed by **Ocean's Eleven** (RePASs: 0.971) across two submissions. These teams demonstrated high entailment, contradiction avoidance, and obligation coverage in their generated answers. Teams **AUEB** and **NLP-MJR** also exhibited strong performance, with RePASs scores of 0.947 and 0.558, respectively.

6 Lessons from RIRAG-2025

The RIRAG-2025 shared task attracted a substantial number of participating teams from both academia and industry. This strong participation underscores the rapid growth and increasing interest in the RegNLP field.

A significant observation during the task was the limited integration of the hierarchical regulatory documents provided into the participants' approaches. Although the teams primarily used the ObliQA dataset, the rich interconnected structure of the entire set of regulatory documents was underutilized. Regulatory rules often refer to or build on one another, and understanding these relationships is crucial for generating accurate and comprehensive answers. Future shared tasks can address this perhaps by providing annotated examples of rule connections and offering detailed guidelines to help participants incorporate these relationships into their system designs.

Table 3: Evaluation Scores of Team Submissions for Subtasks 1 and 2 in the RIRAG-2025 Shared Task, based on a hidden dataset containing 446 questions.

Paper ID	Team Name	R@10	M@10	Es	Cs	OCs	RePASs
Baseline	BM25(passage-only)+GPT-4	0.761	0.624	0.310	0.120	0.176	0.455
Baseline	BM25(rank fusion)+GPT-4	0.764	0.625	0.312	0.125	0.152	0.446
11	USTC-IAT-United *	0.720	0.593	0.777	0.234	0.258	0.600
12	NUST - Group 3 - Team NOVA	0.393	0.227	0.358	0.307	0.109	0.387
13	NUST - Group 1- Team Alpha	0.672	0.521	0.505	0.109	0.098	0.498
14	NUST - Group 2 - Team Omega	0.585	0.097	0.489	0.239	0.167	0.473
15	Havelsan	0.677	0.541	0.330	0.278	0.161	0.404
16	Obayer*	0.780	-	-	-	-	-
17	AICOE	0.633	0.515	**0.827**	0.254	0.230	0.601
18	JurisCore - Submisson 1	0.314	0.093	0.208	0.577	0.005	0.212
	JurisCore - Submisson 2	0.650	0.503	0.395	0.378	0.109	0.375
	JurisCore - Submisson 3	0.650	0.503	0.177	0.716	0.028	0.163
19	Ocean's Eleven - Submission 1	0.686	0.548	**0.986**	0.065	**0.991**	**0.971**
	Ocean's Eleven - Submission 2	**0.694**	0.558	**0.986**	0.062	0.989	**0.971**
	Ocean's Eleven - Submission 3	0.693	0.554	**0.986**	0.149	**0.998**	0.945
20	NLP-MindMappers †	0.662	0.534	0.487	0.174	0.136	0.483
21	NLP-Alpacas * †	0.809	0.625	0.416	**0.046**	0.063	0.477
22	NLP-MJR	**0.731**	**0.602**	0.525	0.156	0.305	0.558
23	TEAM: 1-800	0.705	**0.562**	0.573	0.348	0.090	0.438
24	Indic aiDias - Submission 1	**0.787**	**0.663**	**0.987**	0.062	**0.993**	**0.973**
	Indic aiDias - Submission 2	**0.787**	**0.663**	0.092	**0.037**	0.444	0.316
	Indic aiDias - Submission 3	**0.787**	**0.663**	**0.987**	0.129	0.644	0.834
25	AUEB NLP Group - Submission 1	**0.694**	0.594	0.446	**0.031**	0.502	0.639
	AUEB NLP Group - Submission 2	**0.694**	0.594	0.375	0.110	0.423	0.562
	AUEB NLP Group - Submission 3	**0.694**	0.594	**0.986**	0.096	0.951	**0.947**
26	NLP-LingoLlamas †	0.611	0.499	0.422	0.218	0.048	0.418

Bold values represent the highest performance for each metric.
Teams marked with * could not be evaluated due to incomplete or invalid submissions. Results for these teams are extracted from the original team papers and correspond to evaluations on the ObliQA test set. All other results are based on the hidden dataset of 446 questions.
Teams marked with †did not finalize their camera-ready version for submission.

In the answer generation subtask, we employed RePASs, a metric specifically designed for RIRAG. However, we observed two critical areas for improvement. Firstly, RePASs is currently limited in its ability to evaluate verbatim reproduction of retrieved passages, which can affect the depth and originality of generated answers. Secondly, it lacks a mechanism to evaluate the fluency and cohesion of generated answers. To address these shortcomings, future iterations could enhance RePASs by incorporating penalties for excessive verbatim text and integrating components that assess linguistic quality. Specifically, we will explore the inclusion of semantic similarity thresholds to ensure that generated answers synthesize information rather than directly copying it. Additionally, we intend to incorporate LLM-based evaluations to measure fluency and cohesion, providing qualitative assessments of the generated text.

7 Conclusion

The RIRAG-2025 shared task showcased innovative approaches to tackling the challenges of regulatory information retrieval and answer generation. By leveraging the ObliQA dataset and a robust evaluation framework, participants were able to explore diverse methodologies, from hybrid retrieval systems combining sparse and dense models to advanced generative techniques supported by prompt engineering and post-processing.

While the task brought to light many promising methodologies, it also revealed areas for future exploration. The shared task has set a benchmark for further research in this domain, fostering collaboration between academic and industrial researchers and driving advancements in the automation of regulatory compliance tasks.

Acknowledgments

We would like to express our gratitude to the organizers of COLING 2025 for providing a platform to host the RIRAG-2025 shared task as part of the RegNLP workshop. We thank ADGM for their support in developing the dataset and for expert annotation. We also extend our thanks to all participating teams for their contributions to advancing the field of regulatory information retrieval and answer generation.

References

Tuba Gokhan, Kexin Wang, Iryna Gurevych, and Ted Briscoe. 2024. RIRAG: Regulatory Information Retrieval and Answer Generation. *Preprint*, arXiv:2409.05677.