# A Two-Stage LLM System for Enhanced Regulatory Information Retrieval and Answer Generation

**Fengzhao Sun[1], Jun Yu[1]*, Jiaming Hou[2], Yutong Lin[1], Tianyu Liu[3]**
[1]University of Science and Technology of China,
[2]Harbin Institute of Technology,
[3]Jianghuai Advance Technology Center
sunfz@mail.ustc.edu.cn, harryjun@ustc.edu.cn, 23s105157@stu.hit.edu.cn,
linyutong@mail.ustc.edu.cn, liutianyu18@mails.ucas.ac.cn

## Abstract

This technical report describes our methodology for the Regulatory Information Retrieval and Answer Generation (RIRAG) Shared Task, a component of the RegNLP workshop at COLING 2025. The challenge aims to effectively navigate and extract relevant information from regulatory texts to generate precise, coherent answers for compliance and obligation-related queries. To tackle subtask1, we introduce a two-stage approach comprising an initial output stage and a subsequent refinement stage. Initially, we fine-tune the LLaMa-2-7B model using LoRA to produce a preliminary output. This is followed by the application of an expert mechanism to enhance the results. For subtask2, we design specific prompt to facilitate the generation of high-quality answers. Consequently, our approach has achieved state-of-the-art performance on the leaderboard, which serves as a testament to the effectiveness and competitiveness of our proposed methodology.

## 1 Introduction

Regulatory documents, issued by government bodies, detail compliance rules across various areas like environmental standards and data protection. They are complex, comprehensive, and frequently updated, making them challenging to interpret and keep up with. To manage these documents effectively, specialized NLP techniques, such as information retrieval and question answering, are essential for industries facing governance and compliance challenges.

At the same time, the rapid development of large language models (LLMs)(Brown et al., 2020; OpenAI, 2022; Achiam et al., 2023; Touvron et al., 2023a,b; Chiang et al., 2023) has been remarkable, with major breakthroughs in various fields. This progress implies that LLMs could offer innovative solutions and tools to enhance the processing and comprehension of regulatory documents.

However, simply deploying LLMs in mission-critical domains such as healthcare, law, and finance poses unique challenges that go beyond general AI optimization and alignment. A primary concern is the models' propensity to generate plausible but incorrect "hallucinatory" responses, especially in specialized domains where data is limited or complex. Furthermore, the vast expansion of online data, along with the substantial resources needed for data annotation and model training, makes it difficult for LLMs to stay up-to-date. Recent innovations are trying to tackle these issues. Retrieval-Augmented Generation (RAG)(Lewis et al., 2020) integrates information retrieval to update static knowledge. Chain of Thought(Wei et al., 2022) prompting has led to task-specific workflows. These are enhanced by parameter-efficient fine-tuning methods like Low-Rank Adaptation (LoRA)(Hu et al., 2022) and Reinforcement Learning from Human Feedback (RLHF)(Ouyang et al., 2022), which improve model performance with minimal parameter increases. However, these methods rely on substantial datasets. In response, an Automated Question Passage Generation task for RegNLP has been defined(Gokhan et al., 2024), creating the ObliQA dataset with 27,869 questions from the ADGM financial regulation documents, providing a rich resource for training and refining LLMs in regulatory compliance.

In this study, we introduce a two-stage framework for regulatory document processing: Parameter-Efficient Fine-tuning of Large Language Models (LLMs) and a Hybrid Expert Mechanism. Our key contributions are as follows: (1) Parameter-Efficient Fine-tuning of LLMs: We have fine-tuned the general-purpose LLaMa-2-7B model to specialize in domain-specific retrieval tasks. Our experiments confirm that this approach significantly

*Corresponding author.



*Proceedings of the 31st International Conference on Computational Linguistics*, pages 68–72
January 19–24, 2025. 

enhances the model's ability to accurately retrieve information for regulatory-related queries. (2) Hybrid Expert Mechanism: Beyond conventional rank fusion techniques, we propose a novel expert mechanism designed to refine outputs from various experts. This mechanism ensures a higher level of precision and reliability in the final results.

## 2 Method

### 2.1 Problem Restatement

The Regulatory Information Retrieval and Answer Generation (RIRAG) task encompasses two distinct subtasks:
**Subtask 1**: Passage Retrieval. Given a regulatory question, identify and retrieve the most relevant passages, specifically obligations and related rules, from ADGM regulations and guidance documents.
**Subtask 2**: Answer Generation. Synthesize the retrieved data into precise and informative responses that comprehensively address the regulatory query.

### 2.2 Dataset

In this paper, we mainly use the ObliQA dataset(Gokhan et al., 2024), a multi-document, multi-passage Question Answering (QA) resource designed to advance research in Regulatory Natural Language Processing (RegNLP). The dataset's creation was a three-phase process encompassing Data Collection, Question Generation, and Question-Passages Validation using Natural Language Inference (NLI). Comprising 27,869 questions and their corresponding source passages, ObliQA is sourced entirely from the comprehensive regulatory documentation provided by Abu Dhabi Global Markets (ADGM), the regulatory authority for financial services in the UAE's free economic zones. This dataset is uniquely tailored to facilitate and enhance the development of models capable of retrieving and generating accurate regulatory information.

### 2.3 Passage Retrieval

In this section, we present a practical implementation of the two-stage framework through a case study in the regulatory sector. The overall architecture of our approach is shown in Fig. 1

**Stage 1: Parameter-Efficient Fine-tuning of LLM.** Our first stage is dedicated to specializing the general-purpose LLaMa-2-7B model for regulatory-related retrieval tasks, concurrently mitigating the "hallucination" issues common in LLMs.

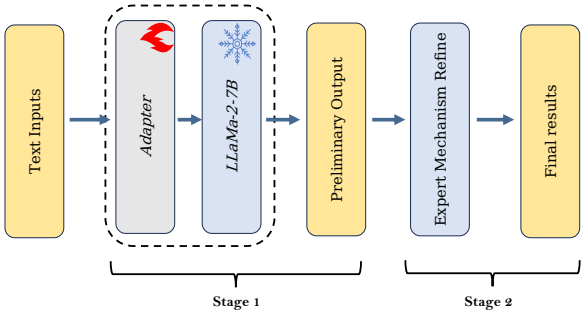


Figure 1: The overall pipeline of our method for Passage Retrieval.

This process aims to yield preliminary results that can be utilized in subsequent stages. During training, we employ the ObliQA dataset to fine-tune the LLaMa-2-7B model. Considering the model's vast parameter count of 7 billion, direct fine-tuning would significantly strain memory resources. To overcome this, we adopt Low-Rank Adaptation (LoRA), an efficient parameter fine-tuning method that allows us to freeze the existing weights and train only a few adapter layers on top of the base model. By adding these adapters to all linear layers of the model, we retain full control over the model's location and network. The following Table 1 provides a detailed overview of our training configuration.

| Hyperparameter | LoRA |
|---|---|
| learning rate | $3 \times 10^{-4}$ |
| batch size | 8 |
| epochs | 50 |
| max length | 128 |
| r | 4 |
| dropout | $1.00 \times 10^{-3}$ |
| alpha | 64 |

Table 1: LoRA hyperparameters.

**Stage 2: Hybrid Expert Mechanism for Refinement** Stage 2 refines Stage 1's results using an expert mechanism that leverages diverse retrieval strengths. It integrates Stage 1's advanced output with traditional methods like BM25, enhancing accuracy and system robustness. The expert mechanism's architecture is depicted in Fig. 2.

This approach offers more flexibility than simple rank fusion by allowing dynamic weight adjustment based on task complexity and expert performance. Spceifically, Let $E_1, E_2, \ldots, E_n$ represent the outputs from the traditional experts (e.g., BM25), and $E_{LLM}$ represent the output from the

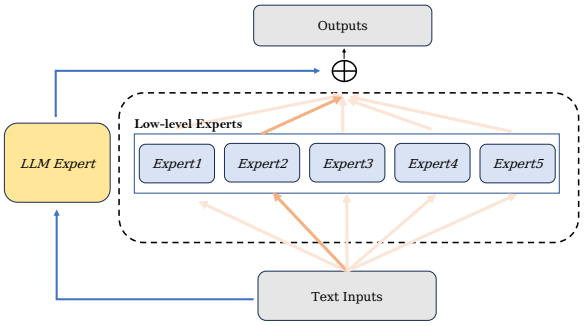


Figure 2: The overall architecture of our Hybrid Expert Mechanism for Refinement.

fine-tuned LLM in Stage 1. Each expert's output is assigned a weight $w_i$ based on its relevance and accuracy. The weighted sum of the expert outputs is calculated as:

$$S = \sum_{i=1}^{n} w_i E_i + w_{LLM} E_{LLM} \quad (1)$$

where $w_{LLM}$ is the weight assigned to the LLM expert.

The output from Stage 1 is considered a high-level expert, and its contribution is combined with the traditional experts' outputs. The final output $Y$ is a weighted combination of all experts' outputs:

$$Y = \frac{\sum_{i=1}^{n} w_i E_i + w_{LLM} E_{LLM}}{\sum_{i=1}^{n} w_i + w_{LLM}} \quad (2)$$

The weights $w_i$ and $w_{LLM}$ can be determined through a training process where the system learns to optimize the combination of experts' outputs for the best performance on a validation set. The combined output $Y$ is then refined through additional processing steps to produce the final result.

We conducted experiments with five retrieval models to serve as our traditional experts: (1) *BM25*(Robertson et al., 1994), a foundational lexical-based model; (2) *DRAGON+*(Lin et al., 2023), a State-of-the-Art (SotA) single-vector dense retriever fine-tuned on the MS MARCO dataset; (3) *SPladev2*(Formal et al., 2021), a SotA neural sparse retriever also fine-tuned on MS MARCO; (4) *ColBERTv2*(Santhanam et al., 2021), a SotA multi-vector dense retriever model, also fine-tuned on MS MARCO. and (5) *Roberta*(Liu et al., 2019), another state-of-the-art method, in our experiments. By integrating the advanced capabilities of the fine-tuned LLM with the robustness of traditional retrieval methods, this hybrid approach leverage the strengths of both to achieve superior performance in regulatory-related retrieval tasks.

### 2.4 Answer Generation

Drawing inspiration from prior work (Gokhan et al., 2024), we initiate the answer generation process once we have identified 10 relevant passages per query from our passage retrieval system. Transitioning to the post-retrieval phase, we apply a scoring-based filtering strategy. This strategy uses a threshold of 0.25 to identify significant drops in relevance between consecutive passages, ensuring a smooth relevance gradient. Furthermore, we enforce a minimum score threshold of 0.7, which ensures that only the most relevant passages are considered for answer generation.

Armed with these carefully selected passages, we utilize the Qwen2-72B model to generate comprehensive answers. The model is guided by a custom prompt designed to simulate the role of a regulatory compliance assistant. This prompt integrates all critical obligations and best practices from the passages into a cohesive response. The prompt is structured as follows:

| **System Prompt** |
|---|
| 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. Please think step by step. |

Table 2: Prompt Design for Regulatory Compliance Assistant.

## 3 Experiment

In this section, we state the details of the experimental implementation. Finally, we present the corresponding experimental results.

### 3.1 Implementation Details

We implement our proposed model using the PyTorch framework. Here are the details of the training process: (1) During the parameter-efficient fine-tuning phase, we train the model with 8 × NVIDIA A100 GPUs. The batch size is set at 8 and adopt the AdamW optimizer with a base learning rate of $3e^{-4}$. Furthermore, we apply the lora technique with a rank of 8 and an alpha value of 64 to fine-tune the model parameters effectively. (2) During hybrid expert mechanism for refinement phase, to

| Model | Passage-only | | Rank fusion | | Hybrid Expert | |
|---|---|---|---|---|---|---|
| | R@10 | M@10 | R@10 | M@10 | R@10 | M@10 |
| BM25 | 64.2 | 50.9 | 64.2 | 51.0 | 64.6 | 51.2 |
| DRAGON+ | 61.4 | 46.3 | 61.9 | 46.3 | 61.5 | 47.7 |
| SPLADE | 64.2 | 49.6 | 64.1 | 49.5 | 63.1 | 55.0 |
| ColBERTv2 | 64.5 | 52.7 | 64.6 | 52.7 | 70.3 | 56.7 |
| ROBERTA | 65.2 | 51.5 | 65.8 | 52.3 | 71.7 | 57.1 |
| LLaMa-7B | 68.4 | 55.4 | 69.0 | 57.0 | 72.0 | 59.3 |

Table 3: Results of the retrieval task on the test dataset. R@10 and M@10 represent Recall@10 and MAP@10, respectively.

| Method | $E_s$ | $C_s$ | $OC_s$ | RePASs |
|---|---|---|---|---|
| BM25(passage-only)+Qwen2 | 0.762 | 0.248 | 0.227 | 0.580 |
| BM25(rank fusion)+Qwen2 | 0.775 | 0.230 | 0.244 | 0.596 |
| BM25(hybrid expert)+Qwen2 | 0.777 | 0.234 | 0.258 | 0.600 |

Table 4: Results of the answer generation task using RePASs on the test dataset. $E_s$, $C_s$, $OC_s$, and RePASs represent Entailment, Contradiction, Obligation Coverage, and RePAS score, respectively.

manage computational efficiency and model input constraints, we truncate both queries and passages to a maximum of 512 tokens.

### 3.2 Metrics

The Passage Retrieval's performance in the RIRAG task is quantitatively assessed through recall@10, thereby enabling the answer-generation module to focus on refining the output. Furthermore, MAP@10 is implemented as a diagnostic tool to evaluate the precision of the ranking within the top-10 retrieved passages. The Answer Generation's performance is evaluated using the RePASs. This metric ensures the answer-generation module produces accurate and consistent responses by integrating three key components: the entailment score ($E_s$), the contradiction score ($C_s$), and Obligation Coverage Score ($OC_s$). The RePASs is calculated as:

$$RePASs = \frac{E_s - C_s + OC_s + 1}{3} \quad (3)$$

### 3.3 Results

The outcomes of the two Sub-Challenges are presented in Table 3 and Table 4, where our method demonstrates superior performance over alternative approaches. The fine-tuning of the LLaMa-7B model proves to be more effective than conventional retrieval techniques. When this is augmented with our hybrid expert mechanism, it creates a synergistic effect that harnesses the collective strengths of a variety of models. This includes those enriched with extensive data training and those with the straightforward efficiency of methods like BM25. Our hybrid approach transcends the limitations of individual experts or simple fusion by incorporating a sophisticated process that refines the output. This process is tailored to the subtleties of the task at hand, resulting in a system that is not only more accurate but also better at generalizing from the data. This comprehensive strategy leads to an enhanced overall performance, making it a more robust solution for the challenges presented by the RIRAG task.

## 4 Conclusion

In summary, this paper delves into the effective utilization of LLMs for regulatory-specific tasks and introduces a hybrid expert mechanism. This innovative approach marries the capabilities of sophisticated LLMs with the tried-and-true methods of traditional retrieval systems. The result is a significant boost in the efficacy of regulatory information retrieval and answer generation processes.

Moving forward, our research will focus on uncovering further optimization techniques for these large models and on broadening their application to a more diverse array of regulatory tasks.

## 5 Limitations

One limitation we encountered is that it does not explore the performance of the model on a broader range of benchmarks. This restraint may limit the

generalizability assessment of the model's applicability to a broader spectrum of downstream tasks, which could be a subject for future work.

## 6 References

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

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

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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 *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

OpenAI. 2022. Chatgpt. https://openai.com/blog/. Chatgpt, 2022. 2, 3, 8.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *Text Retrieval Conference*.

Keshav Santhanam, O. Khattab, Jon Saad-Falcon, Christopher Potts, and Matei A. Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *North American Chapter of the Association for Computational Linguistics*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.