RIRAG: A Bi-Directional Retrieval-Enhanced Framework for Financial Legal QA in ObliQA Shared Task

Xinyan Zhang[1], Xiaobeng Feng*[2], Xiujuan Xu[1], Zhiliang Zheng[3], Kai wu[3],

[1]School of Software Technology,Dalian University of Technology,Dalian,China,
[2]Shanghai University of International Business and Economics,Shanghai,China
,Institute of Change Management and Artificial Intelligence,
[3]School of Computer Science and Technology,Hangzhou Dianzi University,Hangzhou,China,

Correspondence: fxb@suibe.edu.cn

Abstract

In professional financial-legal consulting services, accurately and efficiently retrieving and answering legal questions is crucial. Although some breakthroughs have been made in information retrieval and answer generation, few frameworks have successfully integrated these tasks. Therefore, we propose RIRAG (Retrieval-In-the-loop Response and Answer Generation), a bi-directional retrieval-enhanced framework for financial-legal question answering in ObliQA Shared Task. The system introduces BDD-FinLegal, which means Bi-Directional Dynamic finance-legal, a novel retrieval mechanism specifically designed for financial-legal documents, combining traditional retrieval algorithms with modern neural network methods. Legal answer generation is implemented through large language models retrained on expert-annotated datasets. Our method significantly improves the professionalism and interpretability of the answers while maintaining high retrieval accuracy. Experiments on the ADGM dataset show that the system achieved a significant improvement in the Recall@10 evaluation metric and was recognized by financial legal experts for the accuracy and professionalism of the answer generation. This study provides new ideas for building efficient and reliable question-answering systems in the financial-legal domain. The code of our system is available at https://github.com/Mira-dahu/RIRAG

1 Introduction

Financial-legal question answering systems have emerged as crucial tools for improving access to specialized legal information and services in the financial sector. The complexity of financial-legal documents, combined with the need for accurate and context-aware responses, presents unique challenges in natural language processing. This paper introduces RIRAG, a hybrid system that combines our novel BDD-FinLegal retrieval mechanism, cross-encoding, and advanced language models specifically trained for financial-legal domain question answering.

Recent advances in large language models have revolutionized question answering systems, yet their application in the financial-legal domain remains challenging due to the need for precise citation and adherence to financial regulatory frameworks. Previous approaches have either focused solely on retrieval accuracy or generation quality, often failing to maintain a balance between both aspects. Therefore, we have constructed a completely new system and employed innovative models to address the aforementioned issues. In brief, the contributions of our work are as follows:

- Innovative search mechanism: proposes the BDD-FinLegal dynamic search architecture, which intelligently adjusts traditional and dense embedding methods through query features to achieve more accurate legal document retrieval
- Semantically precise reordering technology: designs a specialized cross-encoder reordering mechanism to significantly improve the relevance and accuracy of legal document retrieval
- Answer generation framework adapted across legal systems: constructs a dual model approach of localization and globalization; achieves comprehensive coverage of legal knowledge in different jurisdictions; and ensures the traceability and professionalism of answers based on expert-annotated datasets

The rest of this paper is structured as follows: Section 2 provides a comprehensive review of existing research on question answering and retrieval systems, identifying key challenges in the domain. Section 3 details our methodology and system

Proceedings of the 31st International Conference on Computational Linguistics, pages 107–113
January 19–24, 2025.

architecture, including the novel BDD-FinLegal mechanism. Section 4 presents the experimental results and analysis. Section 5 discusses the implications and limitations, and Section 6 concludes.

2 Related Work

2.1 Legal Question Answering Systems

Recent advances in natural language processing have yielded sophisticated solutions, moving beyond traditional rule-based systems and keyword matching (Ashley, 2017). Some researchers approach legal QA by utilizing ontologies and knowledge graphs, framing it as an information retrieval challenge (Sovrano et al., 2024). While information retrieval (IR) techniques remain dominant for handling legpal documents and queries (Martinez-Gil, 2023), utilizing large language models represents a promising yet underexplored domain in legal technology.

2.2 Information Retrieval Methods

Dense retrieval has become pivotal in IR with deep neural networks (Luo et al., 2024), demonstrating advantages through continuous vector representations thapt capture semantic relationshipsp. Notable works like DPR (Karpukhin et al., 2020) and ANCE (Xiong et al., 2020) have shown strong performance in open-domain QA tasks. However, applying these methods to legal domains presents unique challenges with terminology, document structure, and citation relationships.

Cross-encopder models have proven effective in reranking initial retrieval results (Nogueira and Cho, 2020), with recent architectures including encoder-decoder and decoder-only models (Déjean et al., 2024). Legal-specific approaches emphasize citation-aware reranking, precedent-based scoring, and hierarchical document structures.

2.3 Hybprid System

Hybrid systems combining multiple components (Zhang et al., 2021) typically employ broad retrieval followed by precise reranking and contextual answer generation. However, current methods often lack context sensitivity and rely heavily on single evaluators familiar with policy corpora (Kalra et al., 2024). Our work builds upon these approaches by introducing novel components specifically designed for legal question answering challenges.

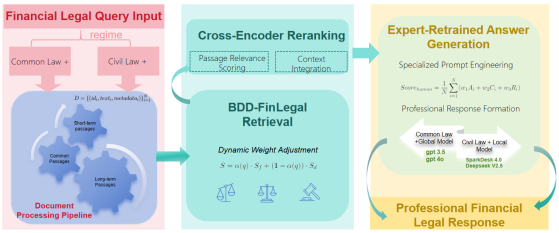

Figure 1: **: An illustration of our system for retrieval-Generation in Legal Question Answering.**

3 Methodology

This section details the architecture and implementation of our RIRAG system, comprising three main components: BDD-FinLegal Retrieval, Cross-Encoder Reranking, and Expert-Retrained Answer Generation.

3.1 System Architecture

The RIRAG system employs a modular architecture designed to handle complex financial legal queries effectively. The system workflow consists of three primary stages:

1. Initial retrieval using the novel BDD-FinLegal approach
2. Reranking of retrieved passages using a specialized cross-encoder
3. Context-aware answer generation leveraging retrained financial-legal expertise

3.2 BDD-FinLegal Retrieval Mechanism

Our novel BDD-FinLegal retrieval mechanism is specifically designed for financial-legal document retrieval:

3.2.1 Dynamic Weight Adjustment

The system implements a sophisticated adaptive weighting scheme:

$$S = \alpha(q) \cdot S_f + (1 - \alpha(q)) \cdot S_d \tag{1}$$

where S_f represents the traditional retrieval score and S_d represents the dense retrieval score.

3.2.2 Adaptive Weighting Scheme

The weight $\alpha(q)$ is dynamically adjusted based on query characteristics:

$$\alpha(q) = \begin{cases} 0.7 & \text{if } |q| < 5 \\ 0.5 & \text{if } 5 \leq |q| < 10 \\ 0.3 & \text{otherwise} \end{cases} \tag{2}$$

3.3 Document Processing Pipeline

The system implements a robust document processing pipeline:

$$D = \{(id_i, text_i, metadata_i)\}_{i=1}^{N} \quad (3)$$

where each document contains:

- Unique identifier (`DocumentID`)
- Passage text
- Passage metadata including `PassageID`

3.4 Expert-Retrained Answer Generation

The answer generation component employs a structured approach with financial legal expertise:

3.4.1 Specialized Prompt Engineering

We implement a domain-specific prompt template as follows: "System: Professional ADGM financial-legal advisor. Guidelines: 1. Base answers on provided financial regulations. 2. Cite specific legal provisions. 3. Use professional financial-legal terminology. 4. Ensure logical completeness. 5. State when information is unavailable."

3.4.2 Context Integration

Retrieved passages are integrated using:

$$C = \sum_{i=1}^{k} w_i \cdot P_i \quad (4)$$

where w_i represents the relevance score and P_i represents the i-th passage.

4 Experiments and Results

This section presents our experimental setup, evaluation metrics, and comparative analysis of different retrieval and generation approaches.

4.1 Experimental Setup

We conducted experiments on the ObliQA dataset[1](Gokhan et al., 2024)

The legal documents included in this dataset cover a range from specific national natural resource assets to current virtual products or services. To address the differences in legal systems across various jurisdictions, we selected two categories of large language models for experimentation: local models and global models. The local models are optimized for Civil Law, while the global models aim to capture the legal principles and applicability of Common Law.

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

4.2 Retrieval Performance Analysis

We compared different retrieval approaches. See Table 1 for the comparison results. We can derive several key insights from the outcomes of our results, our BDD-FinLegal method significantly outperforms traditional approaches across all metrics, achieving a 13.8% improvement in Recall@10 compared to the closest baseline.

Method	R@10	MRR	N@10
TF-IDF	0.456	0.312	0.378
BM25	0.583	0.425	0.491
Dense Retrieval	0.621	0.467	0.535
BDD-FinLegal (Ours)	**0.759**	**0.667**	**0.755**

Table 1: Comparison of Different Retrieval Methods

4.3 Answer Generation Evaluation

Our system was evaluated on the ObliQA Datasets. Table 2 in the appendix shows the results, where Expert-Retrained achieved an overall RePASs score of 0.472, demonstrating the framework's capability in handling financial legal queries. We evaluated multiple language models for answer generation. See Table 3 in the appendix for the evaluation results. From Table 3, we can observe that our Expert-Retrained model demonstrates substantial improvements in both generation quality and accuracy.

4.4 Ablation Studies

We conducted ablation studies to analyze the contribution of each component. The performance difference is calculated as follows. P is the abbreviation of performance:

$$\Delta P = P_f - P_a \quad (5)$$

The ablation study results in Table 4 in Appendix part demonstrate the crucial role of each component in our system's performance. Notably, the removal of the BDD-FinLegal mechanism resulted in the largest performance drop (-8.4%), highlighting its importance in the overall framework.

4.5 Human Expert Evaluation

Legal experts evaluated system outputs based on professional accuracy, citation completeness, and response coherence. Obtain more subjective and nuanced assessment results to help validate the accuracy of automated assessment methods.The score

is calculated as follows:

$$Score_{human} = \frac{1}{N}\sum_{i=1}^{N}(w_1 A_i + w_2 C_i + w_3 R_i) \tag{6}$$

We evaluate using a subset of 40 comprehensive legal documents from the ObliQA dataset, ensuring balanced coverage of domestic-specific regulations and international financial service frameworks. The 40 legal documents are comprehensive enough, ranging from domestic-specific natural resource assets to current virtual products or services. For representative questions randomly selected from the ObliQA dataset, we conduct similarity comparison experiments using Chinese SparkDesk and Deepseek. See Appendix Figure 2 for the results of the experiment.

Where A_i is the accuracy score, C_i is the citation score, R_i is the response coherence score, and w_1, w_2, w_3 are the respective weights. The citation completeness metric directly corresponds to the obligation coverage measure used in RePASs evaluation, providing complementary human validation of our automated metrics.

Detailed evaluation results comparing Global and Local models across different legal systems are presented in Table 5 in the Appendix. The comparison particularly highlights significant differences in handling jurisdiction-specific questions, especially in cases involving financial market infrastructure and liquidation scenarios

5 Discussion

5.1 Key Insights and Implications

Our research provides several significant insights into financial-legal question answering systems:

- The proposed BDD-FinLegal retrieval mechanism demonstrates the effectiveness of dynamically adjusting retrieval strategies based on query characteristics. This approach addresses the inherent variability in financial-legal queries.
- The cross-encoder reranking mechanism significantly enhances the relevance and precision of retrieved passages, a critical aspect in legal document retrieval.
- Expert-retrained language models show substantial improvements in generating contextually accurate and professionally formatted legal responses.

5.2 Limitations

Despite the promising results, our research has several limitations:

The current system faces challenges in data representativeness and potential bias, with restricted generalizability across different legal jurisdictions. Ethical concerns include inherent biases in expert-annotated datasets and the need for robust privacy protection. The dynamic weighting mechanism, while effective, relies on a simple heuristic that requires more sophisticated adaptive strategies. Additionally, the substantial computational resources needed for training and inference may impede widespread deployment.

5.3 Future Research Directions

Future work could focus on:

- Expanding the approach to multi-lingual and cross-jurisdictional legal question answering systems.
- Developing more nuanced adaptive retrieval mechanisms that consider semantic complexity beyond the existing mechanisms..
- Investigating continual learning approaches to keep the system updated with evolving legal frameworks.

6 Conclusion

In this paper, we introduced RIRAG, a bi-directional retrieval-enhanced framework for financial-legal question answering in ObliQA Shared Task. Our key contributions are a dynamic BDD-FinLegal retrieval mechanism adapting strategies based on query characteristics, a specialized cross-encoder reranking approach enhancing passage relevance, and an expert-retrained answer generation framework maintaining high professional standards. Experimental results on the ADGM financial-legal dataset showed significant improvements in retrieval accuracy, answer quality, and expert evaluation metrics, with a Recall@10 of 0.759 and an expert evaluation score of 0.834, outperforming existing approaches.

Our work provides a promising direction for developing more accurate, interpretable, and reliable question-answering systems in the financial-legal domain. By combining advanced retrieval techniques, neural reranking, and domain-specific language models, we have addressed critical challenges in legal information access.

References

Kevin D Ashley. 2017. Artificial intelligence and legal analytics: new tools for law practice in the digital age.

Hervé Déjean, Stéphane Clinchant, and Thibault Formal. 2024. A thorough comparison of cross-encoders and llms for reranking splade. *arXiv preprint arXiv:2403.10407*.

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.

Rishi Kalra, Zekun Wu, Ayesha Gulley, Airlie Hilliard, Xin Guan, Adriano Koshiyama, and Philip Treleaven. 2024. Hypa-rag: A hybrid parameter adaptive retrieval-augmented generation system for ai legal and policy applications. *arXiv preprint arXiv:2409.09046*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Preprint*, arXiv:2004.04906.

Kun Luo, Minghao Qin, Zheng Liu, Shitao Xiao, Jun Zhao, and Kang Liu. 2024. Large language models as foundations for next-gen dense retrieval: A comprehensive empirical assessment. *arXiv preprint arXiv:2408.12194*.

Jorge Martinez-Gil. 2023. A survey on legal question–answering systems. *Computer Science Review*, 48:100552.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert. *Preprint*, arXiv:1901.04085.

Francesco Sovrano, Monica Palmirani, Salvatore Sapienza, and Vittoria Pistone. 2024. Discolqa: zero-shot discourse-based legal question answering on european legislation. *Artificial Intelligence and Law*, pages 1–37.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *Preprint*, arXiv:2007.00808.

Han Zhang, Hongwei Shen, Yiming Qiu, Yunjiang Jiang, Songlin Wang, Sulong Xu, Yun Xiao, Bo Long, and Wen-Yun Yang. 2021. Joint learning of deep retrieval model and product quantization based embedding index. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1718–1722. ACM.

Method	E_s	C_s	OC_s	RePASs
BDD-FinLegal+Deepseek	0.418	0.389	0.387	0.472

Table 2: Results of the answer generation task using RePASs on the evaluation dataset. E_s, C_s, OC_s, and RePASs represent Entailment, Contradiction, Obligation Coverage and RePAS score, respectively.

A Appendix

A.1 Experiment Details

In this section, we will show you some detailed result of experiment.

Model	S_c	L_c	A_c
Base LLM	0.412	0.385	0.723
Fine-tuned LLM	0.445	0.401	0.756
Expert-Retrained (Ours)	**0.502**	**0.458**	**0.834**

Table 3: Comparison of Answer Generation Models

Component	Performance	Relative
Full System	0.834	-
w/o BDD-FinLegal	0.750	-8.4%
w/o Cross-encoder	0.777	-5.7%
w/o Regime-judegment	0.760	-7.4%
w/o Expert-Retrained	0.765	-6.9%

Table 4: Ablation Study Results

A.2 Web interface display

This is a simple web page that we designed for our hybrid-system.

[2]Source: https://adgmen.thomsonreuters.com/rulebook/fund-rules-funds-ver08040723

Q_Index	ModelName	Es	Cs	Ocs	RePASs	Maxs	Mins	AverageScore	AnswerSimilarity
1.0	GPT-4o	0.9615	0.3122	0.3333	0.6609	0.9615	0.3122	0.5670	0.9785
1.0	DeepSeek	0.8875	0.3030	0.1111	0.5652	0.8875	0.1111	0.4667	0.9785
2.0	GPT-4o	0.4624	0.2153	0.3333	0.5268	0.5268	0.2153	0.3845	0.9835
2.0	DeepSeek	0.1712	0.2850	0.3333	0.4065	0.4065	0.1712	0.2990	0.9835
3.0	GPT-4o	0.0536	0.2196	0.0000	0.2780	0.2780	0.0000	0.1378	0.9828
3.0	DeepSeek	0.0737	0.0798	0.3000	0.4313	0.4313	0.0737	0.2212	0.9828
4.0	GPT-4o	0.4357	0.3469	0.8000	0.6296	0.8000	0.3469	0.5530	0.9743
4.0	DeepSeek	0.4963	0.2996	0.4000	0.5322	0.5322	0.2996	0.4320	0.9743
5.0	GPT-4o	0.3253	0.2977	0.3000	0.4425	0.4425	0.2977	0.3414	0.9867
5.0	DeepSeek	0.2791	0.2335	0.2000	0.4152	0.4152	0.2000	0.2820	0.9867
6.0	GPT-4o	0.3572	0.3164	0.4000	0.4803	0.4803	0.3164	0.3885	0.9862
6.0	DeepSeek	0.4046	0.2657	0.7000	0.6130	0.7000	0.2657	0.4958	0.9862
7.0	GPT-4o	0.2427	0.2242	0.1667	0.3951	0.3951	0.1667	0.2572	0.9871
7.0	DeepSeek	0.2018	0.2368	0.0000	0.3217	0.3217	0.0000	0.1901	0.9871
8.0	GPT-4o	0.4495	0.7150	0.3333	0.3560	0.7150	0.3333	0.4634	0.9862
8.0	DeepSeek	0.4988	0.7244	0.4444	0.4063	0.7244	0.4063	0.5185	0.9862
9.0	GPT-4o	0.9806	0.0817	0.6000	0.8330	0.9806	0.0817	0.6238	0.8194
9.0	DeepSeek	0.2440	0.4807	0.0000	0.2544	0.4807	0.0000	0.2448	0.8194
10.0	GPT-4o	0.9886	0.0064	0.8889	0.9570	0.9886	0.0064	0.7102	0.7820
10.0	DeepSeek	0.4230	0.3672	0.3333	0.4630	0.4630	0.3333	0.3967	0.7820

Table 5: Comparative Analysis of Global and Local Models on Legal System-Specific Questions. This table presents a detailed comparison between GPT-4o (representing Global Model + Common Law approach) and DeepSeek (representing Local Model + Civil Law approach) across 10 representative questions from the ObliQA dataset. The evaluation metrics include: Es (Embedding Similarity Score), Cs (Citation Score measuring accurate legal reference usage), Ocs (Obligation Coverage Score), RePASs (Response Professional Accuracy Score), and AnswerSimilarity (similarity score between model outputs). Notable observations: 1) Question 1 demonstrates a clear divergence between Common Law and Civil Law approaches, with GPT-4o showing higher scores across most metrics (Es: 0.96 vs 0.89), reflecting different legal interpretations between the two systems. 2) Questions 9 and 10, which deal with clearing house operations during financial crises, show significant performance gaps. GPT-4o achieves notably higher scores (Es: 0.98, Ocs: 0.89 for Q10) compared to DeepSeek (Es: 0.42, Ocs: 0.33), indicating stronger capabilities in handling complex financial infrastructure scenarios. 3) The overall trend suggests that while both models perform competently, the Global Model (GPT-4o) generally demonstrates more consistent performance across diverse legal contexts, particularly in scenarios requiring cross-jurisdictional understanding. The evaluation was conducted using a subset of 40 comprehensive legal documents, ensuring balanced coverage of both domestic-specific regulations and international financial services frameworks.

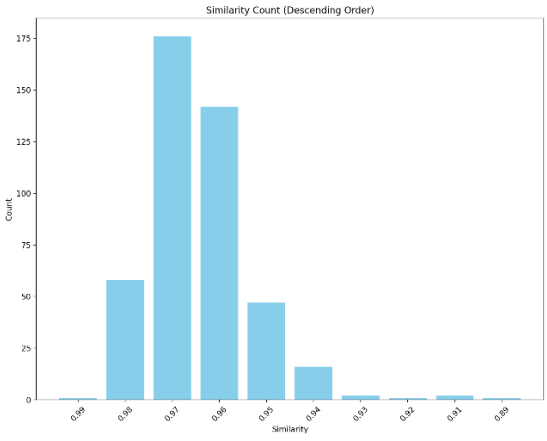

Figure 2: Experimental results of document similarity across legal fields (SparkDesk & Deepseek). We identified the lowest 5 scores corresponding as follows. The lowest scored two queries on virtual assets, followed by the queries on identification of contravention, lastly on definition of a term. The difference might be a results of different jurisdiction applies different legal regimes.[2]

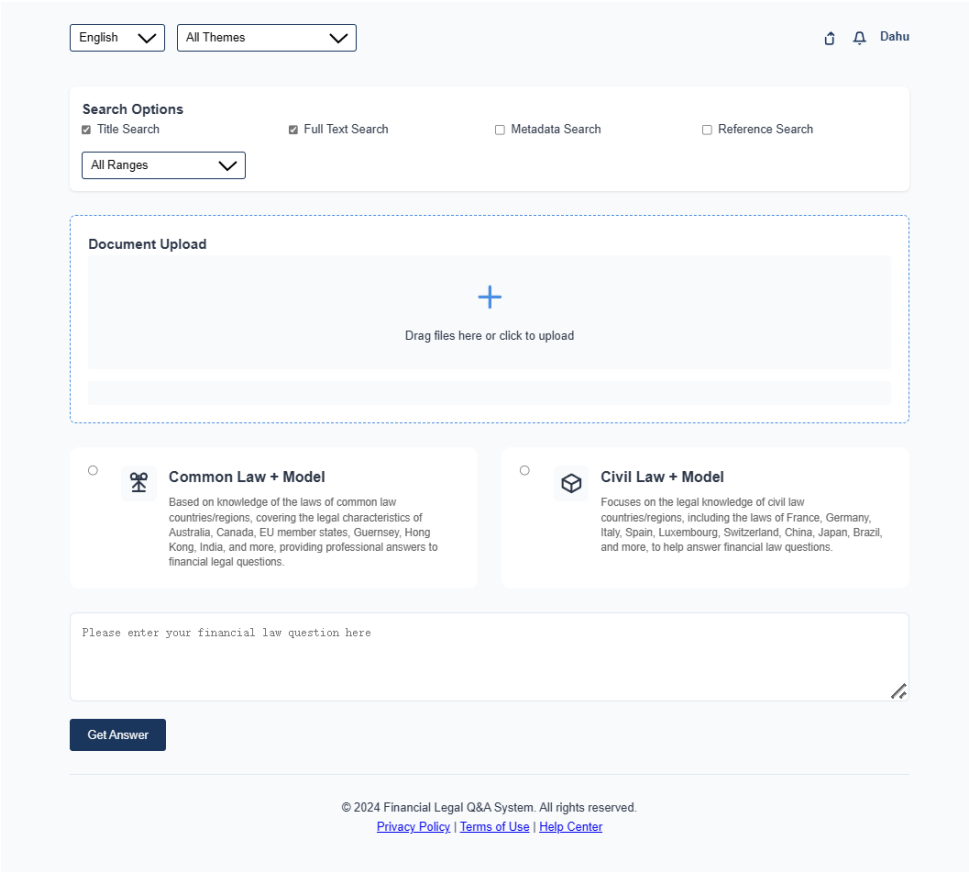

Figure 3: The web interface of the financial law QA system. The interface provides language and topic selection, multiple search options, a document upload function, and the choice of a model based on the common law system and the civil law system. Users can enter questions to get answers. The bottom contains copyright notices and related links.