EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

SEP-LLM: Professional QA in the SEP Domain Using Retrieval-Augmented LLMs

Chenchen Guo[1], Kehao Wang[2], Dianhui Mao[2], Yunlong Xiong[3], Yiwen Lyu[4] and Junhua Chen[1,*]

[1] China National Institute of Standardization, No.4 Zhi Chun Road, Haidian District, Beijing, China
[2] Beijing Technology and Business University, No.11 Fu Cheng Road, Haidian District, Beijing, China
[3] University of Virginia, 1827 University Avenue Charlottesville, VA 22903, United States of America
[4] China Agricultural University, No.17 Qinghuadonglu, Haidian District, Beijing, China

Abstract

INTRODUCTION: Question answering tasks in the Standard Essential Patent (SEP) domain impose high demands on models for professional terminology comprehension, regulatory interpretation, and factual accuracy. Existing general-purpose large language models show limitations in this field, mainly in knowledge retrieval accuracy, semantic matching, and legal compliance of generated content. Therefore, there is an urgent need to develop a specialized intelligent QA system tailored for the SEP domain.
OBJECTIVES: This paper aims to develop an intelligent QA system for the SEP domain, SEP-LLM, to improve knowledge retrieval, semantic matching, and content compliance, providing high-quality automated answers to SEP-related questions.
METHODS: We collected and curated a large set of SEP-related regulations, technical standards, and judicial cases to build a high-quality QA dataset. Leveraging the LightRAG framework, a large language model was used to extract entities and relationships from documents, constructing a structured SEP knowledge graph with incremental updates to ensure dynamic completeness. In retrieval, a two-layer strategy addresses both fine-grained entity queries and broader thematic searches, improving accuracy and coverage. In generation, DeepSeek-LLM-7B was fine-tuned with LoRA on SEP-specific instructions and terminology, enhancing the model's understanding and generation capabilities while significantly reducing training and inference resource requirements.
RESULTS: Experimental results demonstrate that SEP-LLM significantly outperforms leading general-purpose models, including GPT-4o and Qwen3-235B, across three key metrics: BLEU-4, ROUGE-L, and Accuracy. These findings underscore its superior performance and promising potential for professional Quality Assurance within SEP domain.
CONCLUSION: The LightRAG-based SEP-LLM system effectively enhances knowledge retrieval, semantic understanding, and compliance in SEP QA tasks, demonstrating the potential of retrieval-augmented generation techniques in specialized domains and providing a practical solution for intelligent information services in the SEP field.

Keywords: Standard Essential Patent(SEP), Retrieval-Augmented Generation(RAG), Knowledge Graph, Low-Rank Adaptation (LoRA), Instruction Fine-Tuning

Received on 20 October 2025, accepted on 20 March 2025, published on 11 May 2026

doi: 10.4108/eetsis.10653

1. Introduction

Standard Essential Patent (SEP), as a key innovative resource arising from the close integration of standards and patents, has become a strategic priority highly regarded by multiple countries and international organizations worldwide.[1] The regulatory and policy framework related to SEP is extensive and dispersed, encompassing laws, departmental regulations, judicial decisions, and technical standards, and is continually evolving. This poses

*Corresponding author. Email: chenjunh@cnis.ac.cn

dual challenges for policymakers and market participants in terms of information integration and compliance assessment. In the context of an increasingly complex international environment, traditional reliance on manual processing and supervision has proven insufficient for efficient and precise regulatory interpretation.[2]

Meanwhile, with the rapid advancement of Large Language Model (LLM) technologies, artificial intelligence (AI) has demonstrated significant improvements in natural language understanding and generation tasks. Several studies have explored its applications in the legal domain. For instance, Song et al.[3] proposed a multi-label classification method for legal documents that combines domain-specific pretraining with a label-attention mechanism, achieving superior performance on datasets such as POSTURE50K compared to multiple baseline systems, demonstrating the potential of deep learning in legal text classification. Wang et al.[4] developed a BERT-based legal question answering system that leverages vector representations and the Milvus retrieval engine to efficiently match and respond to Chinese legal queries, validating the feasibility of AI-powered question answering in specialized domains. Liga et al.[5] fine-tuned GPT-3 on GDPR-related regulations for legal rule classification, achieving results that significantly outperform prior experiments, indicating that LLMs can effectively identify legal clauses and norms even in low-resource settings. These studies provide a solid technical foundation for tasks including legal text comprehension, question answering, and rule classification.

However, conventional LLMs generally rely on static knowledge acquired during training, and their closed knowledge bases cannot guarantee the timeliness and authoritativeness of regulations in the SEP domain.[6] For example, after the release of new regulations, the lack of an immediate update mechanism may lead to outdated or incorrect outputs. Moreover, general-purpose LLMs (e.g., ChatGPT, Claude) often perform inadequately when confronted with SEP texts that are dense with specialized terminology and tightly structured logic, limiting their ability to provide accurate interpretations. To overcome these limitations, Retrieval-Augmented Generation (RAG) has emerged as an effective solution. RAG combines an LLM generator with an external knowledge base, retrieving relevant document passages corresponding to a query and integrating them into the prompt context to enhance both the accuracy and timeliness of generated answers.[7]

Despite RAG's substantial extension of knowledge coverage and improvement in generation quality, its practical application in the SEP domain still faces several challenges. First, SEP regulatory data is highly heterogeneous, including SEP statements, laws, regulations, guidelines, judicial cases, technical standards etc., with multi-layered semantic structures and rich implicit relationships. Traditional unstructured knowledge bases struggle to represent and retrieve this information efficiently. Second, retrieval methods based on static semantic similarity are insufficient to handle the polysemy, specialized terminology, and complex logical relations typical of texts in SEP domain, limiting both accuracy and coverage. To address these issues, several advanced RAG methods have been propose4d in recent years. Zhao et al.[8] introduced LongRAG, which employs a dual-perspective modeling mechanism to capture both global context and local factual details, alleviating the "lost in the middle" problem in long-context multi-hop question answering and achieving a 17.25% improvement over Vanilla RAG. Gokdemir et al.[9] developed HiPerRAG, leveraging high-performance computing (HPC) and multimodal document parsing to index and retrieve over 3.6 million scientific articles, improving the accuracy of scientific QA. Lim et al.[10] proposed MacRAG, which constructs multi-scale adaptive contexts by compressing and hierarchically retrieving documents and progressively merging relevant contexts, optimizing long-text handling. Jin et al.[11] presented LongRefiner, which applies document structure and multi-task learning for adaptive filtering of redundant information. However, these methods generally involve high computational costs and complex system architectures, limiting their large-scale deployment in SEP scenarios.

To address the aforementioned challenges, this paper proposes a SEP-focused intelligent question answering system based on LightRAG, aiming to manage SEP policy knowledge and provide QA functionality with high efficiency and relatively low cost. At its core is an LLM specifically trained for the SEP domain. By constructing a structured knowledge graph and employing a dual-layer retrieval strategy, LightRAG not only efficiently manages massive and heterogeneous SEP regulatory data but also precisely captures diverse user query intents—ranging from fine-grained entities to macro-level thematic concepts. Furthermore, its incremental update mechanism ensures the real-time currency of the knowledge base, providing an ideal technical framework for building a high-performance and accurate intelligent QA system for SEP. Drawing on the LightRAG framework, the system constructs a structured knowledge graph to systematically represent related laws, regulations, guidelines, cases and standards, enabling efficient integration of multi-source regulatory data. The retrieval module adopts a dual-layer mechanism, combining entity-based local semantic search with relation-based global semantic recall, improving alignment between query intent and structured semantics. In the generation phase, the system fuses structured semantic representations with original text context, enhancing logical coherence and cross-paragraph integration of answers. Furthermore, Low-Rank Adaptation (LoRA) is employed for instruction fine-tuning of the LLM, improving its comprehension of SEP-specific terminology and regulatory logic while effectively controlling resource consumption during fine-tuning and deployment. The workflow of SEP-LLM is shown in Figure 1.

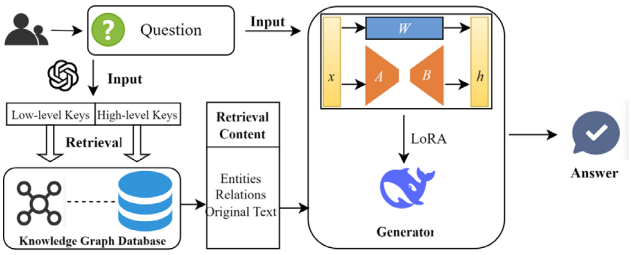

Figure 1. SEP-LLM System Workflow

2. Materials and Methods

2.1. DataSet

SEP data is characterized by its high dispersion, structural complexity, and domain-specific nature.[12]Currently, SEP-related information is widely scattered across various international organizations, national regulations, and industry standards, lacking a unified organizational structure and standardized format. This leads to fragmented data resources that are difficult to systematically integrate and utilize. Such limitations significantly constrain the knowledge coverage and response effectiveness of large language model (LLM)-based intelligent question-answering systems in the SEP domain.[13]

To address these challenges, this study systematically integrates 89 SEP-related regulatory and policy documents from multiple countries and international organizations worldwide. This document collection spans multiple dimensions, comprehensively covering the entire process of SEP regulation and various stakeholders, fully reflecting the diversity of regulatory content, geographical origins, and formal structures.

First, the documents cover three critical phases of SEP regulation: pre-event, in-process, and post-event. Among them, there are 46 pre-event documents, primarily involving early-stage policies such as standard development and patent applications; 10 in-process documents, covering procedural regulations such as examination, licensing, and dispute resolution; and 33 post-event documents, including enforcement, supervision, and dispute settlement measures. This full-lifecycle coverage not only demonstrates the systematic and diverse nature of SEP regulatory policies but also provides rich contextual and policy references for intelligent question-answering systems.

Second, the document sources encompass multiple countries and international organizations, including China, the United States, and the International Organization for Standardization (ISO), ensuring broad geographical and institutional representation. By incorporating regulations from leading countries and authoritative international organizations in the SEP field, this study enhances the authoritativeness and general applicability of the dataset, thereby improving the model's adaptability across different jurisdictional environments. The quantitative distribution of specific source organizations/countries is illustrated in the corresponding figure 2.

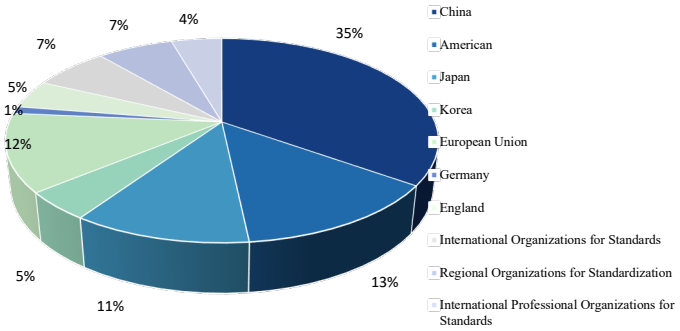

Figure 2. Distribution of SEP Documents by Organization/Country

Furthermore, this study ensures diversity in document types, including departmental regulations, local standards, operational guidelines, acts, laws, and more. The content spans multiple levels, such as regulatory formulation, implementation details, judicial interpretations, and policy guidance, fully reflecting the complexity and hierarchical nature of the SEP regulatory system. This lays a solid foundation for the fine-grained structuring of the knowledge base and the instruction tuning of the question-answering model. The quantitative distribution of specific document types is shown in the corresponding figure 3.

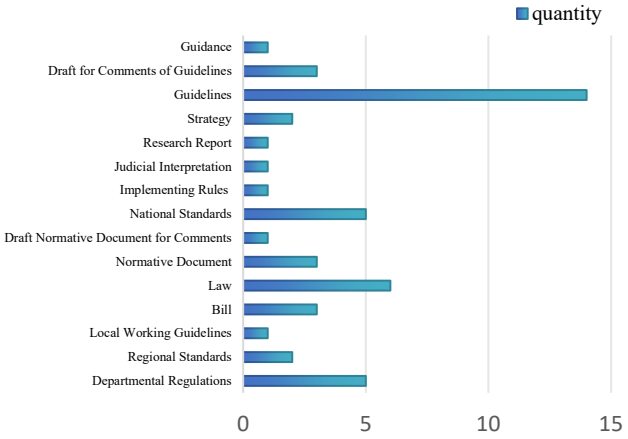

Figure 3. Distribution of SEP Document Types by Quantity

Based on the above SEP regulatory documents, this study carries out two core tasks to fully exploit and utilize the professional value of the data:

1.For the knowledge base module in the Retrieval-Augmented Generation (RAG) system, the LightRAG framework is employed to convert unstructured regulatory documents into a structured knowledge graph, enhancing semantic expressiveness and retrieval efficiency.

2.Leveraging large language models for information extraction, combined with manual cleaning and validation, to construct a high-quality SEP question-answering dataset for subsequent instruction tuning and system evaluation.

Knowledge Graph for Standard Essential Patents

In the SEP domain, the design of the regulatory policy knowledge base directly affects both the traceability of regulatory information and the accuracy of intelligent question answering. Traditional Retrieval-Augmented Generation (RAG) systems typically segment regulatory texts into fixed-size chunks and index them using vector embeddings. However, such chunk-based storage often disrupts the logical connections and contextual coherence between regulatory provisions, leading to fragmented retrieval results that fail to fully capture the requirements of the regulations. Moreover, SEP regulation involves multi-level and multi-dimensional information, including patent holders, standards organizations, technical standards, and legal provisions.[14] A text-chunk-based knowledge base struggles to effectively model these complex relationships, resulting in insufficient information association and limiting the performance of QA systems.

To overcome these limitations, this paper adopts the LightRAG data processing approach[15], leveraging a knowledge graph structure to construct the knowledge base, thereby enhancing the accuracy and reliability of RAG systems in regulatory QA tasks within the SEP field. A Knowledge Graph explicitly represents the core entities and their relationships in the SEP domain, preserving contextual information from regulatory texts while establishing a complete regulatory chain. For example, a knowledge graph can link a patent with its related technical standard, the involved enterprise, and the corresponding legal judgment, enabling the QA system to quickly aggregate multi-dimensional information and improve both the depth of regulatory understanding and semantic association.

Specifically, the process begins with chunking SEP regulatory and policy documents, dividing the original texts into relatively independent, manageable paragraphs to improve efficiency in subsequent extraction and retrieval. Next, a Pretrained Language Model is applied to perform entity recognition and relation extraction, identifying key entities in the regulations (e.g., patents, standards, enterprises, legal provisions) and their interconnections. For instance, the model can capture relations such as “a patent belongs to an enterprise” or “a regulation cites a technical standard,” thereby forming a semantic network of entities and relations. After deduplication and structural optimization, the extracted entity nodes and relational edges are stored in a graph database, resulting in a structured knowledge graph. Meanwhile, vector representations are generated for each entity and its associated textual content using embedding techniques, constructing an efficient vector index to support fast retrieval. A sample subgraph extracted from the knowledge graph is illustrated in the figure 4.

To ensure the integrity of the knowledge graph construction, we implemented a multi-stage quality control workflow. First, we performed entity and relation extraction by using a DeepSeek-LLM-7B model fine-tuned on domain-specific data, which demonstrated high precision in recognizing SEP-domain terminology during internal testing. Second, the extracted results underwent automated de-duplication and consistency validation. Finally, we conducted a manual sampling audit where domain experts evaluated 100 randomly selected entity-relation triplets. The expert review confirmed that over 90% of the triplets accurately reflected the semantics of the source documents, with only minor instances requiring merging or correction. This rigorous pipeline ensures that the constructed knowledge graph possesses the high accuracy and completeness necessary to provide a reliable knowledge foundation for subsequent retrieval and QA tasks.

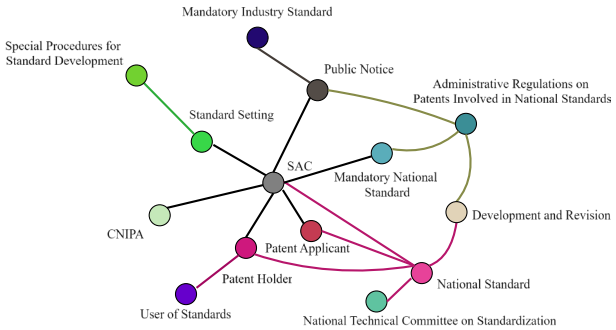

Figure 4. Example of a SEP Knowledge Graph (Partial View)

Question Answering Dataset for the SEP Domain

To optimize the generator module and enhance its performance in SEP-oriented QA tasks, this study constructed a multi-source, domain-specific SEP QA dataset based on the aforementioned documents.[16] The construction process is as follows: Large Language Models (LLMs) were employed to extract “instruction–input–output” triplets from the original regulatory texts, conforming to the linguistic characteristics of legal and regulatory language (an example structure is illustrated in Figure 5) [17] .To ensure that the generated content not only adheres to the formal conventions of regulatory language but also achieves comprehensive semantic coverage and sufficient reasoning depth, multi-level task instructions were incorporated into the prompt design. These included diverse case-based QA scenarios such as semantic interpretation of individual provisions and applicability analysis under contextual backgrounds. In addition, by alternately constructing samples across different regulatory types (e.g., technical specifications, administrative measures, penalty rules), the dataset was enriched with diverse QA pairs and improved applicability in real-world use cases[18].

After the generation phase, multiple rounds of manual review and revision were conducted by domain experts and annotators[19]. The review process focused on removing redundant or semantically irrelevant questions, correcting factual errors in the answers, refining the expressions to align with the rigor and professionalism of regulatory texts, and strictly ensuring logical consistency and semantic coherence among the instruction, input, and output. The final dataset not only provides high-quality, scenario-oriented training samples for LoRA-based instruction tuning but also serves as a reliable benchmark dataset for evaluating model performance in regulatory intelligent QA tasks.

"1":{
"instruction":"Does a transaction involving standard-essential patents constitute a concentration of undertakings, and how should it be assessed?",

"input":"Based on the Anti-Monopoly Law and relevant regulations, explain the standards and considerations for determining whether a transaction involving Standard Essential Patents (SEPs) constitutes a concentration of undertakings.",

"output":"The determination of whether a transaction involving standard-essential patents (SEPs) constitutes a concentration of undertakings shall be made based on a comprehensive analysis in accordance with the Anti-Monopoly Law, the Guidelines of the State Council's Anti-Monopoly Committee on the Anti-Monopoly Review in the Field of Intellectual Property, and relevant anti-monopoly regulations. At the same time, two key factors should be given particular consideration: first, whether the products or services covered by the SEPs constitute an independent business unit, or whether they can generate independent and quantifiable revenue; and second, the specific terms and duration of the SEP licenses. If these factors indicate that the transaction constitutes a concentration of undertakings and meets the notification thresholds specified in the Measures of the State Council on the Notification Thresholds for Concentrations of Undertakings, the relevant undertakings are obligated to submit a notification to the competent anti-monopoly enforcement authority before implementing the concentration. Failure to do so renders the implementation of the concentration prohibited."
}

Figure 5. SEP QA Dataset Example

2.2. LightRAG

LightRAG is an innovative Retrieval-Augmented Generation (RAG) framework specifically designed for complex and heterogeneous knowledge environments. By constructing a structured knowledge graph and incorporating a unique dual-layer retrieval algorithm, LightRAG enhances the accuracy and efficiency of large language models (LLMs) in domain-specific question answering. In the SEP domain, where regulatory texts, judicial decisions, and licensing agreements are characterized by complex structures and frequent updates, LightRAG provides robust technical support for intelligent regulatory Q&A through fine-grained indexing, flexible retrieval, and efficient generation mechanisms.

Specifically, in the indexing stage, LightRAG segments texts into finer-grained chunks to facilitate efficient localization and management. The system first utilizes the Generator LLM to parse user queries through a specialized prompt, guiding it to output 'high-level' concepts (e.g., organizations, domains) and 'low-level' entities (e.g., temporal data, formats, essential elements) relevant to the query intent. Subsequently, a dual-pronged retrieval approach is executed: high-level keywords are employed for global semantic recall within the knowledge graph to aggregate relevant thematic subgraphs, while low-level keywords are used for precise local semantic retrieval within the vector index to pinpoint the most relevant text segments. For instance, when processing a lengthy patent licensing agreement, the document can be split into clause-level segments, enabling subsequent queries to focus on individual provisions. Then, with the help of LLMs, key entities within the text—such as patent holders, licensing scope, and validity period—along with their legal relationships, are automatically identified and organized into a knowledge-graph-based structured index. For judicial decisions, LightRAG further deduplicates and merges identical legal entities and reasoning across different cases, reducing redundancy. Meanwhile, LLM-driven analytic functions generate key–value pairs for each entity and relation, where the key represents the entity name or keyword, and the value provides a corresponding text summary, ensuring efficient matching in later retrieval. In addition, LightRAG supports incremental update algorithms that seamlessly integrate newly issued regulatory documents into the existing graph, ensuring the timeliness and completeness of the knowledge base.

In the retrieval stage, LightRAG employs a dual-layer retrieval mechanism tailored to both detail-oriented queries (e.g., specific regulatory clauses) and abstract queries (e.g., overall SEP policy trends or case law analysis). For example, when querying the legal obligations of a particular licensing agreement, low-level retrieval pinpoints the relevant clauses and legal entities with high precision; whereas for broader inquiries such as "Evolution of Negotiation Rules for Global SEP Licenses " high-level retrieval aggregates entities and relations across multiple policies and case documents, offering comprehensive trend insights. During retrieval, LLMs first extract both global and local keywords from the query, which are then matched against corresponding entities and relations in the vector database and knowledge graph. Leveraging the multi-hop connection capability of the graph, LightRAG expands the retrieval scope by incorporating relevant subgraphs and neighboring nodes, effectively covering complex regulatory logic and multidimensional semantic associations. This significantly enhances the system's ability to respond to complex queries in the SEP domain.

Finally, in the generation stage, LightRAG integrates the retrieved structured knowledge graph information with corresponding textual content as input to the LLM. This hybrid approach avoids the bias of relying solely on unstructured text, ensuring that the generated answers are both comprehensive and domain-appropriate. For instance, in SEP regulatory Q&A, the model can synthesize structured legal clause information with unstructured judicial case descriptions to produce answers that are not only aligned with legal norms but also adapted to practical application scenarios[15]. The detailed technical workflow of combining structured and unstructured information is illustrated in the figure 6.

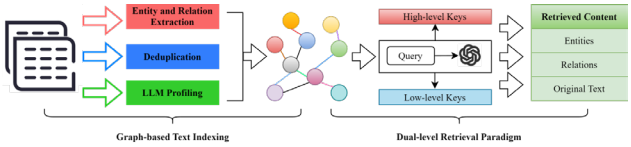

Figure 6. LightRAG Technology Roadmap

2.3. Low-Rank Adaptation for Instruction Fine-Tuning

In a typical RAG architecture, the generator is usually based on a general-purpose Large Language Model (LLM), such as DeepSeek, Qwen, or ChatGPT. These models are trained on massive amounts of unstructured, domain-agnostic corpora and exhibit strong language understanding and generation capabilities. However, due to the lack of training on domain-specific knowledge and task patterns, such models encounter a series of application bottlenecks in the SEP domain[20]. On one hand, their instruction-following capability is limited, making it difficult to precisely execute complex queries with fixed logical structures, such as legal clause interpretation or patent-standard correlation evaluation. On the other hand, these models often struggle with output format control, frequently producing redundant content and unstructured expressions, which severely undermines their practicality and reliability in professional Q&A tasks.

To address these issues, Instruction Fine-Tuning has been proposed[21]. Compared with traditional fine-tuning approaches, instruction fine-tuning not only emphasizes the injection of domain knowledge but also focuses on enhancing the model's adaptability to task-specific intentions. Concretely, by constructing structured triplet-form datasets (instruction–input–output), the model can be trained to (i) strengthen its ability to parse professional query templates (e.g., "Does Patent X comply with Article Z of Standard Y?"), and (ii) improve its capacity to generate outputs that capture upstream–downstream logical chains, legal terminology boundaries, and formatting conventions within SEP scenarios, thereby ensuring that responses are lawful, coherent, and practically useful.

Nevertheless, applying instruction fine-tuning with a full-parameter approach requires optimizing all parameters of the model[22], which incurs prohibitive computational costs and demands substantial hardware resources. Moreover, as SEP regulations and standards continue to evolve and update frequently, models that are overly fitted to specific instruction templates may struggle to adapt to newly issued rules or standards, thus impairing their generalization ability. To overcome these challenges, Low-Rank Adaptation (LoRA) [23]is introduced for instruction fine-tuning, enabling the model to enhance rule understanding and generation capabilities with significantly reduced computational overhead. LoRA is an efficient parameter fine-tuning method whose core idea is to decompose the weight matrix of the Pretrained Language Model (PLM) into two low-rank matrices, thereby constraining the degrees of freedom in weight updates and reducing computational costs.

Specifically, let the weight matrix of a certain neural network layer be $W \in \mathbb{R}^{d\times k}$, which represents the inherent knowledge acquired during pre-training from general-purpose corpora, including language ability and common sense. Since pre-training data is broad but not tailored for the SEP domain, the model often performs poorly on tasks related to standard-essential patents.

In traditional fine-tuning, the entire parameter matrix W must be updated to adapt the model to a specific task. However, this approach involves a massive number of parameters, resulting in high computational and storage costs. To address this, LoRA (Low-Rank Adaptation) introduces a low-rank update term on top of the frozen weights without directly modifying W. The formulation is:

$$W' = W + \Delta W \tag{1.}$$

where ΔW denotes the task-specific update for the SEP domain. Instead of optimizing a full-rank matrix, LoRA approximates ΔW as the product of two low-rank matrices:

$$\Delta W = BA \tag{2.}$$

Here, $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$, with rank $r \ll min(d,k)$,which drastically reduces the degrees of freedom for parameter updates. Intuitively, ΔW serves as a "patch" that injects domain-specific knowledge such as patent clauses, licensing conditions, and legal terminology. During training, the original weight matrix W remains frozen, while only the parameters of A and B are optimized. Thus, the forward propagation can be reformulated as:

$$h = Wx + \Delta Wx = Wx + BAx \tag{3.}$$

where x is the input vector (e.g., the embedding of SEP-related text). Here, Wx encodes general linguistic knowledge, while BAx adds the SEP-specific knowledge update.

To ensure stable training, LoRA adopts special initialization and scaling strategies:

$$A \sim \aleph(0, \sigma^2), B = 0 \tag{4.}$$

This ensures that $\Delta W = 0$ at the initial state, preventing interference with the base model. Furthermore, the update term is scaled during training:

$$\Delta W = \frac{\alpha}{r} BA \tag{5.}$$

Where α is a scaling factor that balances the update magnitude. With this design, LoRA enables large language models to quickly adapt to SEP-specific knowledge at very low computational cost, while significantly enhancing their performance on complex regulatory and legal reasoning tasks.

During the LoRA fine-tuning process, the optimal hyperparameters were determined via a grid search. Specifically, we set the rank of the low-rank matrices to r=8. We utilized a learning rate of 2×10^{-4}, a batch size of 4, and trained the model for 3 epochs. These parameters yielded the best performance on our validation set.

3. Results

3.1. Experimental Setup

To validate the effectiveness of the proposed method in the Question Answering (QA) task within the domain of SEP, we designed a series of experiments. The experiments were conducted on the SEP QA dataset we constructed, which was randomly split into training and testing sets at an 8:2 ratio. During the evaluation phase, the model received natural language queries from the test set and generated corresponding answers through the generation module. Its performance was quantified by comparing the generated outputs with human-annotated gold-standard answers.

In terms of evaluation metrics, given that QA tasks in the SEP domain are characterized by highly structured professional expressions and strict requirements for terminological accuracy, we adopted three complementary indicators—BLEU-4, ROUGE-L, and Accuracy—covering dimensions of generation quality, semantic coverage, and factual correctness.

BLEU-4, which measures n-gram (here 4-gram) precision, was used to assess the consistency between generated answers and reference answers in terms of linguistic expression and terminology reproduction. This metric emphasizes precise word matching and sequence accuracy, which is critical for evaluating whether the model can accurately restate key legal terms or technical phrases in statutes and patent descriptions, thereby ensuring the professionalism and rigor of generated content. ROUGE-L, based on the longest common subsequence (LCS), was employed to evaluate the semantic overlap between generated and reference answers. Unlike BLEU-4, it places more emphasis on content coverage and logical coherence rather than strict word order. In the SEP context, this helps determine whether the model fully covers core elements such as patent claims, legal provisions, or technical solutions, avoiding the omission of crucial information. Accuracy, on the other hand, measures the exact match between the generated output and the reference answer, serving as an indicator of reliability in fact reasoning and causal inference. For example, for queries such as "If Patent A is revoked, is Patent B still valid?", which require precise logical reasoning, Accuracy directly reflects the correctness of the model's inference, thereby providing a clear basis for evaluating its performance in complex reasoning tasks.

Regarding the experimental environment, model training and inference were performed on a single NVIDIA GeForce RTX 4090 GPU. With 24 GB of memory and excellent single-card computing capacity, the RTX 4090 supports the complete Retrieval-Augmented Generation (RAG) workflow while significantly reducing deployment and experimental costs compared to data center GPUs such as A100 or H800. Without the need for multi-GPU parallelism, it enables efficient adaptation of SEP-LLM, demonstrating strong resource utilization and practical deployability. The experimental pipeline was built on the PyTorch deep learning framework and integrated with the Hugging Face ecosystem, facilitating future model extensions and rapid deployment in domain-specific scenarios.

3.2. Generator Comparison and Selection

In the Retrieval-Augmented Generation (RAG) framework, the performance of the generator determines the system's ability to understand retrieved information and to organize coherent answers. Therefore, selecting a generator with strong instruction-following ability, knowledge integration capacity, and high-quality language generation is critical to the overall performance of the system.

In our experiments, we conducted a comparative evaluation of four mainstream open-source Chinese large language models—ChatGLM, DeepSeek, Baichuan, and Qwen—on the SEP-focused question answering dataset we constructed. Using the test set described in Section 3.1, we calculated three evaluation metrics for each model: BLEU-4, ROUGE-L, and Accuracy. The results are presented in Table 1.

Table 1. Performance Comparison of Candidate Generator Models on the SEP QA Dataset

model	Evaluation Metrics		
	BLEU-4	ROUGE-L	Accuracy(%)
Baichuan2-7B-Chat	4.41	10.99	29.10
GLM-4-9B-Chat	6.39	14.47	31.85
Qwen2-7B-Chat	6.41	14.29	33.03
deepseek-llm-7b-Chat	**8.95**	**19.01**	**35.44**

The results show that DeepSeek-LLM-7B-Chat achieved superior performance across all three metrics, with BLEU-4 (8.95), ROUGE-L (19.01), and Accuracy (35.44). This indicates that DeepSeek-LLM-7B-Chat possesses stronger capabilities in understanding retrieved content, extracting key information, and generating answers in the standardized legal language style required in the SEP domain. For example, when answering the question regarding the "timing requirement for SEP declaration," DeepSeek-LLM-7B-Chat not only correctly stated that most standard-setting organizations encourage patent holders to disclose their SEPs early in the standard-setting process., but also added the crucial legal consequence that "failure to declare in time may result in the loss of the patent holders' right to assert exclusivity based on the SEP." The response was articulated in a rigorous and legally precise manner.

By contrast, the other models demonstrated performance gaps relative to DeepSeek-LLM-7B-Chat. Qwen2-7B-

Chat and GLM-4-9B-Chat showed comparable results on BLEU-4 (6.39 and 6.41, respectively) and ROUGE-L (14.47 and 14.29, respectively). However, in terms of Accuracy, Qwen2-7B-Chat performed better (33.03), outperforming GLM-4-9B-Chat (31.86) by 1.17 percentage points. This suggests that Qwen2-7B-Chat exhibits slightly greater stability in instruction execution under certain complex QA scenarios. On the other hand, Baichuan2-7B-Chat ranked lowest overall, particularly in BLEU-4 (5.12) and Accuracy (27.45), indicating weaknesses in fine-grained information extraction and adaptation to the specialized linguistic style of the SEP domain.

Therefore, based on the comprehensive evaluation, this study ultimately selects DeepSeek-LLM-7B-Chat as the generator to ensure optimal generation quality and stability in SEP-oriented question answering tasks.

3.3. Analysis of the LoRA Fine-Tuning Training Loss

In the LoRA-based instruction fine-tuning experiments, we adopted the Cross-Entropy Loss function to measure the difference between the model's output and the target text. The loss is formally defined as:

$$Loss = \sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t}\, log\, y'_{i,t} \qquad (6.)$$

where N denotes the number of samples in a batch, T represents the sequence length, $y_{i,t}$ is the one-hot distribution of the target token, and $y'_{i,t}$ is the predicted probability distribution. This loss function calculates the token-level difference between the prediction and the ground truth, and minimizes this divergence so that the model's output progressively approximates the target text.

Figure 7 illustrates the training loss curve of the Qwen-7B model during the fine-tuning process, including both the raw loss values and the smoothed trend line. It can be observed that the loss decreases rapidly in the early stage of training, indicating that the model quickly learns core concepts in the SEP domain, such as the roles of international standardization bodies like ETSI and 3GPP, as well as key terms including the FRAND (Fair, Reasonable, and Non-Discriminatory) licensing principle and SEP declaration procedures.

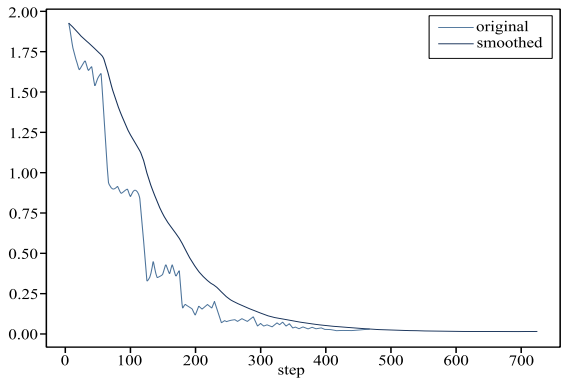

Figure 7. Training Loss Curve

As training progresses, the rate of loss reduction slows and eventually converges, suggesting that the model transitions into a stage of fine-grained optimization and knowledge integration. At this stage, the focus shifts toward enhancing the ability to understand complex instructions and adapting to the linguistic style of the domain. For instance, when handling instructions such as "summarize the situations involving patents as stipulated in the specific standards." the model is able to accurately capture the task intent and generate responses aligned with patent law and standardization practices, rather than providing overly generic answers. In terms of language style, the model gradually acquires professional expressions consistent with formal documents such as policies, regulations, as well as the explanatory articles of professional institutions.

Overall, the training loss curve is smooth and converges stably without significant oscillations, demonstrating that the fine-tuning process is both effective and controllable. These results indicate substantial improvements in terminology recognition, semantic understanding, and standardized expression, enabling the model to better adapt to specialized question answering tasks in the SEP domain.

3.4. Performance Analysis

To evaluate the domain-specific performance of SEP-LLM in SEP tasks, this study selected several representative general-purpose large language models (LLMs) from both domestic and international institutions as baselines, including Qwen3-235B, GLM-4.5, DeepSeek-V3, LLaMA 3.1, and GPT-4o. These models are flagship LLMs released by various organizations, demonstrating strong performance in general natural language processing tasks and being widely validated across multiple benchmark datasets. The purpose of this experiment is to verify whether SEP-LLM, optimized for the SEP domain, can achieve more accurate and efficient performance in this vertical domain compared with these state-of-the-art general models. The experimental results are presented in Table 2.

Table 2. Performance Comparison of Candidate Generator Models on the SEP QA Dataset

model	Evaluation Metrics		
	BLEU-4	ROUGE-L	Accuracy(%)
Qwen3-235b	17.82	28.65	41.20
GLM-4.5	16.47	27.18	41.03
Deepseek-V3	18.02	29.14	43.19
LLaMA 3.1	15.36	26.07	38.98
GPT-4o	19.45	30.92	44.00

SEP-LLM	**29.67**	**39.41**	**49.88**

The results show that in SEP-related question answering tasks, SEP-LLM significantly outperforms the general-purpose models across all evaluation metrics. Specifically, its BLEU-4 score reaches 29.67, which is higher than GLM-4.5, LLaMA 3.1, and Qwen3-235B by 13.20, 14.31, and 11.85, respectively. This demonstrates that SEP-LLM is capable of generating professional expressions more closely aligned with SEP terminology, legal provisions, and technical details. For example, when answering queries related to SEP disclosure, SEP-LLM can accurately cite clauses from standardization organizations, integrate relevant technical implementation details, and produce high-quality outputs consistent with industry norms—greatly enhancing its practical value in professional scenarios.

In terms of semantic matching, SEP-LLM achieves a ROUGE-L score of 39.41, which is 8.49 higher than the second-best model GPT-4o (30.92). This indicates that SEP-LLM can more precisely capture semantic associations among legal concepts and technical contexts, thereby producing more reliable and well-organized answers.

For question-answering accuracy, SEP-LLM attains 49.88%, reflecting excellent factual consistency and instruction-following capabilities. This is particularly crucial in highly sensitive tasks such as SEP infringement determination and standard correlation analysis, ensuring both the compliance and interpretability of generated content.

It is worth noting that the general-purpose models (e.g., DeepSeek-V3 and GPT-4o), supported by massive training data and powerful computational resources, remain leaders in general NLP scenarios. However, in the highly specialized and regulation-intensive SEP domain, SEP-LLM demonstrates significant advantages through domain-specific optimizations—such as the integration of structured SEP knowledge, the incorporation of retrieval-augmented generation (RAG), and instruction fine-tuning aligned with legal linguistic styles. These results validate the effectiveness of domain-targeted optimization strategies and highlight that lightweight models, when carefully designed for specialized domains, can rival or even surpass general-purpose LLMs in vertical professional applications.

This integrated advantage is empirically validated by the subsequent ablation study results (see Table 3). Specifically, removing the LightRAG module results in a 10.50% decrease in accuracy, underscoring the vital importance of knowledge enhancement for factual precision. Furthermore, the removal of LoRA instruction fine-tuning leads to a significant decline in BLEU-4 and ROUGE-L scores, highlighting its critical role in improving linguistic fluency and semantic coverage. The synergy between these two components collectively drives the superior performance of SEP-LLM.

3.5. Ablation Study

To verify the contribution of each component to model performance, we conducted ablation experiments and evaluated the model under four different configurations. Specifically, we compared the following setups: the original deepseek-llm-7b-Chat model (w/o), the model without the LightRAG module (w/o LightRAG), the model without the LoRA instruction fine-tuning module (w/o LoRA), and the full model SEP-LLM. The experimental results are shown in Table 3. It should be clarified that "w/o LightRAG" refers to "using traditional text-based vector retrieval (Naive RAG) as the retrieval module, while the generator remains the LoRA-fine-tuned DeepSeek-LLM-7B."

Table 3. Ablation Study Results of SEP-LLM

model	Evaluation Metrics		
	BLEU-4	ROUGE-L	Accuracy(%)
w/o (Baseline)	8.95	19.01	35.44
w/o LightRAG (Naive RAG)	22.88	32.73	39.38
w/o LoRA (LightRAG)	15.77	29.39	46.59
SEP-LLM (Full)	**29.67**	**39.41**	**49.88**

The results demonstrate that SEP-LLM, benefiting from both the LightRAG and LoRA instruction fine-tuning modules, achieves the best scores across all three metrics—BLEU-4, ROUGE-L, and Accuracy—reaching 29.67, 39.41, and 49.88, respectively. Compared to the w/o configuration (deepseek-llm-7b-Chat), these represent improvements of 20.72, 20.40, and 14.44, respectively. This indicates that the model not only generates high-quality text well aligned with the standard-essential patent (SEP) domain, but also ensures factual reliability and legal compliance in its outputs.

When removing the LoRA module (w/o LoRA), BLEU-4 and ROUGE-L drop significantly to 15.77 and 29.39, while Accuracy decreases only slightly to 46.59 (a reduction of 3.29). This suggests that LoRA instruction fine-tuning primarily enables the model to better interpret user instructions and generate text more consistent with user intent by leveraging pre-defined answer templates, thereby yielding substantial improvements in fluency and semantic coverage. However, its effect on factual consistency and precise question answering is limited, which still relies heavily on external knowledge augmentation to ensure accuracy and verifiability.

When removing the LightRAG module (w/o LightRAG), BLEU-4, ROUGE-L, and Accuracy drop to 22.88, 32.73, and 39.38, respectively, representing reductions of 6.79, 6.68, and 10.50. These results highlight the significant role of LightRAG in improving both generation quality and

question-answering accuracy. Its retrieval-augmented mechanism effectively compensates for the model's knowledge gaps in the SEP domain, resulting in more stable performance in professional question-answering tasks. For instance, in queries concerning SEP determination criteria, the absence of LightRAG often leads to incomplete clause citations or incorrect dates, whereas the full model
accurately references the corresponding regulations and provisions.
Finally, when all modules are removed, the original model (w/o) performs at the lowest level across all three metrics, with BLEU-4, ROUGE-L, and Accuracy of only 8.95, 19.01, and 35.44, respectively. This underscores that without domain knowledge augmentation and instruction fine-tuning, large language models struggle to generate high-quality text for SEP-specific tasks.

3.6. Case Analysis

This section presents a concrete question-answering case in the SEP domain (as shown in the table) to systematically compare the differences between traditional Retrieval-Augmented Generation (NaiveRAG) and SEP-LLM in terms of retrieval, knowledge integration, and answer generation. The case focuses on the issue of patent disclosure in the ETSI standard-setting process, providing an intuitive illustration of the significant differences between the two methods in both workflow and final output. As shown in Table 4:

Table 4. Comparative Examples of Domain Question Answering

Original Query	During the ETSI standard-setting process, when and in what form should patent holders disclose patents that may potentially become SEPs? What elements should be included in the disclosure?	
Method	NaiveRAG	LightRAG
Retrieval Strategy	Single-layer vector retrieval	Double-layer retrieval
Retrieved Content	**A**: Members and participants shall disclose to the ETSI Secretariat any patent or patent application that they believe may become essential to ETSI's work, as early as reasonably practicable. The disclosure should include information such as the patent or application number, filing date, patent holder, relevant technical clauses, or links to the relevant draft standard. **B**: If a patent is still in the application stage, the patent holder should submit the declaration before the relevant technology is formally adopted into the standard. Once a disclosure declaration is submitted, the patent holder shall update it promptly in case of changes in patent status or ownership. ETSI will record the disclosure in its IPR database and make it publicly accessible.	**Keywords:** **High-level:** ETSI, SEP (Standard Essential Patent), Patent Holder **Low-level:** Timing of Disclosure, Form of Disclosure, Elements of Disclosure **Retrieved Triplets (Partial):** ① ETSI Policy, Definition, Timing of Disclosure ② Patent Holder, Through, ETSI IPR Information Statement and Licensing Declaration Form... **Retrieved Content Fragments (Partial):** "Members and participants shall disclose to the ETSI Secretariat any patent or patent application that they believe may become essential to ETSI's work, as early as reasonably practicable. The disclosure shall include information such as the patent or application number, filing date, patent holder, relevant technical clauses, or links to the relevant draft standard."
Response Content	According to the ETSI Intellectual Property Rights (IPR) Policy, patent holders should disclose to the Secretariat any patents that may become Standard Essential Patents (SEPs) as early as possible. The disclosure must include the patent number, application number, filing date, patent holder information, and relevant technical details. Furthermore, ETSI encourages the use of its online declaration form for submissions and requires timely updates to the disclosure information when patent status changes, for recording in the ETSI IPR database.	According to the ETSI IPR Policy, patent holders should respond to the Chair's IPR call at the beginning of each technical meeting and disclose patents that may eventually become SEPs as early as possible. Disclosure must be made online through the dedicated form provided by ETSI, although signed emails are also accepted. The disclosure must include certain mandatory information (such as patent numbers, application numbers, filing dates, patent holders, and relevant technical sections) and should include a commitment to license on FRAND terms.

In the traditional RAG pipeline, the system vectorizes the user query and retrieves several of the most similar long text chunks from the document repository. Although these chunks are related to ETSI's IPR policy, they are lengthy

and unstructured. For example, the retrieved texts A, B, and C not only contain large amounts of context irrelevant to the core question, but also mix together key information such as “disclosure timing,” “disclosure form,” and “disclosure elements.” These unstructured results are directly fed into the large language model (LLM) as context, forcing the model to spend additional computational resources to identify, extract, and integrate effective information, thereby increasing the reasoning burden. As a result, although the generated answer is semantically correct to some extent, it lacks clear structure, which undermines both credibility and professionalism.

By contrast, SEP-LLM leverages a unique combination of knowledge graph integration and multi-granularity retrieval, substantially improving the quality of the retrieved content. The user query is first parsed into high-level and low-level keywords, enabling precise capture of user intent. Based on these keywords, the system retrieves concise entities, relational triples, and highly relevant text fragments from the knowledge graph, which are then concatenated to form the context. This retrieved content is highly structured, explicitly linking knowledge points that were originally scattered across different texts—such as ETSI policies, definitions, disclosure timing, and disclosure elements—into a coherent framework. Such structured context not only reduces the number of tokens the LLM needs to process, but also greatly enhances reasoning efficiency and answer accuracy.

Ultimately, the LightRAG-enhanced model generates answers that clearly address questions such as “when,” “in what form,” and “which elements” in a structured manner, while also providing traceability to specific sources in the knowledge base. This high-quality response not only offers users more accurate and professional solutions but also significantly improves the explainability and reliability of the entire QA system.

4. Discussion

This study successfully constructs a high-performance intelligent question answering system for the SEP domain by organically integrating LightRAG and LoRA instruction tuning technologies. Comparative experimental results demonstrate that within this highly specialized vertical domain, the SEP-LLM model significantly outperforms general-purpose large language models (LLMs) across BLEU-4, ROUGE-L, and accuracy metrics, fully affirming the effectiveness of knowledge enhancement and instruction tuning strategies tailored for specialized legal-technical fields.

Ablation studies indicate that the superior performance of SEP-LLM under low-cost conditions primarily stems from the synergistic effects of two technical pathways: enhanced retrieval-augmented generation and LoRA-based instruction tuning. LightRAG effectively compensates for the lack of domain-specific information in the language model's training data, strengthening its knowledge retention and logical reasoning capabilities. Meanwhile, LoRA instruction tuning enhances the model's ability to respond accurately to domain-specific task demands and adhere to formal expression standards. This combined strategy of "knowledge enhancement + instruction tuning" offers a new technical paradigm for intelligent question answering in the legal and SEP domains, with strong potential for generalization.

Specifically, the LightRAG module in SEP-LLM constructs a dynamically updatable structured knowledge graph through fine-grained document segmentation and entity-relationship extraction based on LLMs. This mechanism not only enables a systematic representation of the multi-layered relationships among complex regulatory provisions, judicial precedents, and technical standards in the SEP domain but also ensures the conciseness and timeliness of the knowledge base under low-cost conditions through graph deduplication and incremental updates. The two-tier retrieval strategy—entity-level retrieval for precise information localization and relation-level retrieval for aggregating macro-thematic information—effectively improves both the breadth of retrieval recall and the accuracy of matching. This allows the model to flexibly mobilize knowledge resources according to different query intents. These designs significantly enhance the model's robustness when confronted with professional semantic ambiguities, polysemous terms, and logically complex questions. Compared to existing question answering systems based on unstructured text retrieval, LightRAG demonstrates stronger adaptability and efficiency in cross-document multi-hop reasoning and knowledge fusion, solidifying the core value of structured knowledge representation in the intelligent parsing of legal-technical content.

At the generator level, LoRA instruction tuning further promotes the model's in-depth understanding of the professional language style and instruction semantics in the SEP domain. During fine-tuning, the model not only learns a multitude of core SEP concepts and professional terminology but also masters the rigorous expressive norms and logical structures of legal provisions. This results in generated responses that are not only linguistically fluent but also compliant with legal industry standards. Although fine-tuning contributes limited direct improvement to factual accuracy (which relies primarily on the retrieval enhancement module), it plays a key role in enhancing textual semantic coverage, logical consistency, and expressive normativity, highlighting the unique value of instruction tuning in optimizing fine-grained control and domain adaptation of language models.

It is important to note that while SEP-LLM achieves certain performance improvements in domain-specific question answering tasks, there remains considerable room for enhancement in the evaluation metrics. This phenomenon is closely related to the inherent complexity of SEP domain tasks: firstly, SEP documentation is highly specialized and legally rigorous, involving numerous cross-references between technical features and legal clauses, which places extremely high demands on answer precision; secondly, questions often require multi-hop reasoning and cross-

document knowledge integration, necessitating that the model simultaneously comprehends technical details and legal connotations, significantly increasing task difficulty; furthermore, the evaluation criteria are exceptionally strict—even if the main body of an answer is correct, inaccuracies in detailed expression or incomplete citation of provisions lead to penalties. These factors collectively contribute to a less dramatic performance gain compared to some closed-domain tasks (where accuracy for some simple QA tasks can exceed 80%). Future work may explore finer-grained knowledge injection mechanisms, more powerful multi-hop reasoning architectures, and stricter domain-adaptive training strategies to continually improve the system's performance in complex professional domains.

5. Conclusion

This study systematically explores the application of retrieval-augmented generation (RAG) and LoRA-based instruction tuning in intelligent question answering within the highly specialized domain of SEP, aiming to address challenges such as fragmented information, complex knowledge structures, and insufficient reliability in professional QA tasks. The main contributions are threefold:

1. To address the scarcity of high-quality data in the SEP domain, we constructed a multi-source dataset covering policies and regulations, standards documents, representative cases, and practical updates, and generated high-quality instruction-tuning QA pairs based on this dataset, providing a valuable foundational resource for RAG research in the SEP field.
2. To enhance the model's ability to integrate and associate specialized knowledge, we incorporated the LightRAG framework, which significantly improves performance in complex regulatory parsing and multi-hop reasoning tasks through structured knowledge indexing and semantic retrieval optimization.
3. Leveraging LoRA instruction tuning, we achieved deep domain adaptation of the generative model at low cost, enabling it to better understand the SEP task context and generate responses that are both legally rigorous and technically accurate.

Experimental results demonstrate that the proposed method significantly outperforms baseline models in terms of accuracy, knowledge utilization efficiency, and robustness, validating the feasibility and effectiveness of combining LightRAG with LoRA instruction tuning in vertical professional domains. Ablation studies and case analyses further confirm the core value of LightRAG in knowledge enhancement and cross-document reasoning, as well as the critical role of LoRA instruction tuning in task instruction comprehension and adaptation to domain-specific language norms.

Despite these achievements, the study has certain limitations. For instance, the system's coverage of cross-lingual SEP policies remains limited, and information may be missing when processing regulations from different countries or regions. The model's reasoning capability and answer completeness for extremely complex cases or multi-hop reasoning tasks still require improvement. Additionally, the integration of multi-source information and the dynamic updating of the knowledge base need further optimization. Future work will explore multi-modal knowledge injection, dynamic knowledge base updates, and interpretable reasoning mechanisms to further enhance the system's applicability, reliability, and explainability in real-world industrial scenarios, providing a feasible technical solution for intelligent question answering in the SEP domain and broader specialized vertical fields.

Acknowledgments

This article is sponsored by the Basic Research Projects of CNIS "Research on the Necessity Assessment Method for Patents related to standards (Number of the project: 572025Y-12476)" and "Research on Key Methods and Paths for the Integration of Patents and Standards (Number of the project: 572024Y-11408)".

References

[1] YANG Z, WU X. Measurement of information disclosure level in standard-essential patent databases[J]. China Invention & Patents, 2025, 22(1): 4-15.
[2] KANG S. Challenges in the Determination of Standard-Essential Patents[EB]. (2024-01-02).
[3] SONG D, VOLD A, MADAN K, et al. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training[J]. Information Systems, 2022, 106: 101718.
[4] WANG C, LUO X. A Legal Question Answering System Based on BERT[C]. Proceedings of the 2021 5th International Conference on Computer Science and Artificial Intelligence. New York, NY, USA: Association for Computing Machinery, 2022: 278-283.
[5] LIGA D, ROBALDO L. Fine-tuning GPT-3 for legal rule classification[J]. Computer Law & Security Review, 2023, 51: 105864.
[6] Mansurova A, Mansurova A, Nugumanova A. QA-RAG: Exploring LLM reliance on external knowledge[J]. Big Data and Cognitive Computing, 2024, 8(9): 115.
[7] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2020: 9459-9474.
[8] ZHAO Q, WANG R, CEN Y, et al. LongRAG: A Dual-Perspective Retrieval-Augmented Generation Paradigm for Long-Context Question Answering[C]. AL-ONAIZAN Y, BANSAL M, CHEN Y N. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics, 2024: 22600-22632.
[9] GOKDEMIR O, SIEBENSCHUH C, BRACE A, et al. HiPerRAG: High-Performance Retrieval Augmented Generation for Scientific Insights[C]. Proceedings of the Platform for Advanced Scientific Computing Conference. 2025: 1-13.

[10] LIM W, LI Z, KIM G, et al. MacRAG: Compress, Slice, and Scale-up for Multi-Scale Adaptive Context RAG[A]. arXiv, 2025.
[11] JIN J, LI X, DONG G, et al. Hierarchical Document Refinement for Long-context Retrieval-augmented Generation[A]. arXiv, 2025.
[12] PENTHEROUDAKIS C. TECHNICAL AND PRACTICAL ASPECTS RELATED TO PATENT QUALITY IN THE CONTEXT OF STANDARD ESSENTIAL PATENTS-An exploratory case study for WIPO[J].
[13] BARON J, POHLMANN T. Mapping standards to patents using declarations of standard-essential patents[J]. Journal of Economics & Management Strategy, 2018, 27(3): 504-534.
[14] TYAGI A, CHOPRA S. Standard Essential Patents (SEP's)-Issues & Challenges in Developing Economies[J]. Journal of Intellectual Property Rights, 2017, 22: 121-135.
[15] GUO Z, XIA L, YU Y, et al. LightRAG: Simple and Fast Retrieval-Augmented Generation[A]. arXiv, 2025.
[16] NAZAR W, NAZAR G, KAMIŃSKA A, et al. How to Design, Create, and Evaluate an Instruction-Tuning Dataset for Large Language Model Training in Health Care: Tutorial From a Clinical Perspective[J]. Journal of Medical Internet Research, 2025, 27: e70481.
[17] MA Y, MIZUKI S, FUJII K, et al. Building Instruction-Tuning Datasets from Human-Written Instructions with Open-Weight Large Language Models[A]. arXiv, 2025. DOI:10.48550/arXiv.2503.23714.
[18] ABDALLA M, KASEM M S, MAHMOUD M, et al. ReceiptQA: A Question-Answering Dataset for Receipt Understanding[J]. Mathematics, 2025, 13(11).
[19] LIU Q, NIU Z, LIU S, et al. iTRI-QA: a Toolset for Customized Question-Answer Dataset Generation Using Language Models for Enhanced Scientific Research[A]. arXiv, 2025.
[20] YANG H, ZHANG Y, XU J, et al. Unveiling the Generalization Power of Fine-Tuned Large Language Models[A]. arXiv, 2024.
[21] CHUNG H W, HOU L, LONGPRE S, et al. Scaling instruction-finetuned language models[J]. Journal of Machine Learning Research, 2024, 25(70): 1-53.
[22] KUMAR A, RAGHUNATHAN A, JONES R, et al. Fine-tuning can distort pretrained features and underperform out-of-distribution[J]. arXiv preprint arXiv:2202.10054, 2022.
[23] HU E J, SHEN Y, WALLIS P, et al. Lora: Low-rank adaptation of large language models.[J]. ICLR, 2022, 1(2): 3.