applied sciences

MDPI

Article

Enhanced Retrieval-Augmented Generation Using Low-Rank Adaptation

Yein Choi [1], Sungwoo Kim [2], Yipene Cedric Francois Bassole [2] and Yunsick Sung [2,*]

[1] Computer Science and Engineering (CSE), Dongguk University-Seoul, Seoul 04620, Republic of Korea; pt840072@dgu.ac.kr
[2] Department of Computer Science and Artificial Intelligence, Dongguk University-Seoul, Seoul 04620, Republic of Korea; sean1255@dgu.ac.kr (S.K.); 2024126796@dgu.ac.kr (Y.C.F.B.)
* Correspondence: sung@dongguk.edu

Abstract: Recent advancements in retrieval-augmented generation (RAG) have substantially enhanced the efficiency of information retrieval. However, traditional RAG-based systems still encounter challenges, such as high latency in output decision making, the inaccurate retrieval of road traffic-related laws and regulations, and considerable processing overhead in large-scale searches. This study presents an innovative application of RAG technology for processing road traffic-related laws and regulations, particularly in the context of unmanned systems like autonomous driving. Our approach integrates embedding generation using a LoRA-enhanced BERT-based uncased model and an optimized retrieval strategy that combines maximal marginal similarity score thresholding with contextual compression retrieval. The proposed system enhances and achieves improved retrieval accuracy while reducing processing overhead. Leveraging road traffic-related regulatory datasets, the LoRA-enhanced model demonstrated remarkable performance gains over traditional RAG methods. Specifically, our model reduced the number of trainable parameters by 13.6% and lowered computational costs by 18.7%. Performance evaluations using BLEU, CIDEr, and SPICE scores revealed a 4.36% increase in BLEU-4, a 6.83% improvement in CIDEr, and a 5.46% improved in SPICE, confirming greater structural accuracy in regulatory text generation. Additionally, our method achieved an 8.5% improvement in retrieval accuracy across key metrics, outperforming baseline RAG systems. These contributions pave the way for more efficient and reliable traffic regulation processing, enabling better decision making in autonomous systems.

Keywords: Retrieval-Augmented Generation (RAG); Low-Rank Adaptation (LoRA); road traffic legal Information Retrieval

Academic Editor: Luis Javier Garcia Villalba

Received: 13 March 2025
Revised: 7 April 2025
Accepted: 15 April 2025
Published: 17 April 2025

Citation: Choi, Y.; Kim, S.; Bassole, Y.C.F.; Sung, Y. Enhanced Retrieval-Augmented Generation Using Low-Rank Adaptation. *Appl. Sci.* **2025**, *15*, 4425. https://doi.org/10.3390/app15084425

1. Introduction

Recent research has focused on applying retrieval augmented generation (RAG) [1] technology in this domain, particularly for traffic law documents, to efficiently retrieve complex road traffic-related laws and regulations. RAG is an advanced system that simultaneously performs information retrieval and text generation, enabling the swift and accurate processing of road traffic-related laws composed of specialized terminology and various regulations. In unmanned systems, especially autonomous driving [2,3], compliance using road traffic-related laws and regulations from a specific jurisdiction is essential.

However, the application of RAG-based road traffic-related laws and regulations retrieval in this domain presents several challenges that must be addressed.

First, latency in decision making can be a significant issue. RAG involves both information retrieval and text generation, the latency introduced can hinder timely responses to rapidly changing road conditions, such as sudden pedestrian crossings or unexpected obstacles. Traditional RAG-based systems operate through a sequential process where document retrieval and embedding computation must fully complete before text generation can begin. This two-stage architecture creates significant computational bottlenecks, as the system must process and rank large volumes of road traffic-related laws and regulations before generating responses for decision making. In autonomous driving scenarios where milliseconds matter, these processing delays may compromise the safety and reliability of decision-making systems, particularly when navigating complex traffic situations requiring immediate action.

Another challenge is the potential inaccuracy in retrieving road traffic-related laws and regulation information. RAG system may sometimes retrieve outdated or irrelevant traffic regulations that do not align with current road conditions. This inaccuracy stems from the complexity of road traffic-related laws and their specialized terminology, establishing a direct causal link to non-compliance. When retrieval systems extract inappropriate road traffic-related standards, autonomous systems make decisions based on incorrect regulatory frameworks. This misalignment propagates through the decision-making process, causing the system to implement actions that violate current local traffic laws. Thus, retrieval accuracy directly determines whether an autonomous vehicle operates within road traffic boundaries, creating a clear causal relationship between information retrieval quality and regulatory compliance.

Finally, computational overhead and scalability constraints pose additional challenges. simultaneously addressing complex road scenarios requires extensive retrieval and text generation computations. This high computational burden not only increases operational costs but also creates bottlenecks that limit scalability and reduce accuracy in decision making, particularly in resource-constrained autonomous driving environments.

To improve the retrieval and processing of road traffic-related laws and regulations, it is essential to address key challenges inherent in conventional retrieval and representation methods. Traditional retrieval methods suffer from multiple limitations, including issues with semantic representation, contextual loss, and relevance assessment.

One key issue is the difficulty in representing the semantic meaning of road traffic-related laws and regulation texts with traditional embedding techniques. Pre-trained language models often fail to capture the subtle nuances and specialized terminology found in road traffic-related laws and regulations, such as those in the “Road Traffic Act”. This incomplete semantic representation makes it difficult to distinguish between highly relevant and less relevant road traffic-related laws and regulation information.

Another limitation is the loss of context during document preprocessing and chunking. Standard preprocessing techniques segment road traffic-related laws, regulations, and documents into fixed token units, which risks fragmenting critical road traffic legal clauses or disrupting contextual continuity. This context loss reduces the precision and reliability of retrieved road traffic-related laws and regulation information.

Lastly, traditional retrieval strategies often fail to accurately assess relevance. Conventional retrieval methods, such as FAISS-based vector search combined with maximal marginal relevance (MMR) [4] and similarity score thresholding, may not adequately account for the unique characteristics of road traffic-related laws and regulations. As a result, these methods often retrieve outdated or irrelevant information that does not reflect the latest traffic regulations or regional specifics, potentially leading to flawed decision making in unmanned systems.

Given these challenges and gaps in existing approaches, the main objectives of this research are to enhance the semantic representation capabilities of retrieval models for road traffic-related laws and regulations; preserve contextual continuity during document processing; and improve the accuracy and relevance of retrieved road traffic legal information.

In this paper, we propose a low-rank adaptation (LoRA) [5]-enhanced RAG framework to improve the accuracy and efficiency of road traffic-related law and regulation retrieval. This framework addresses the above challenges through an optimized retrieval mechanism.

Our method builds on the established evidence of LoRA effectiveness in adapting language models while using significantly fewer parameters. Prior research has demonstrated that LoRA can considerably reduce the parameter count while achieving good performance. For BERT-based models specifically, studies have shown that LoRA adaptations can improve domain-specific retrieval capabilities while reducing computational requirements [6].

The proposed approach consists of three key components:

First, we introduce a LoRA-enhanced bidirectional encoder representations from transformers BERT-based uncased retriever. We refine the retrieval process by applying LoRA to the upper transformer layers, improving the model's ability to capture complex road traffic-related laws and regulation information while reducing computational overhead.

Second, we employ a multi-stage filtering mechanism. The retrieval system leverages FAISS-based vector search to obtain relevant road traffic-related laws and regulations and refines the results through a combination of MMR and transformer attention-based filtering [7].

Third, we integrate context-augmented road traffic law text generation. Those retrieved documents are integrated into the prompt of a language model to ensure that generated responses adhere to road traffic-related laws and regulations and road traffic legal standards.

By combining these components, the proposed method enhances road traffic-related laws and regulations retrieval, supporting more reliable decision making in unmanned systems.

This paper makes the following key contributions:

- We introduce a LoRA-enhanced RAG framework to improve retrieval efficiency. By applying LoRA to a BERT-based uncased retriever, we enhance its ability to capture complex road traffic legal contexts while reducing the processing overhead.
- We introduce a multi-stage retrieval filtering mechanism [3] to refine the selection of relevant road traffic-related laws and regulations. By combining FAISS-based retrieval with MMR and transformer attention-based [8] filtering, we improve both the accuracy and diversity of retrieved road traffic-related laws and regulations.
- We integrate context-augmented query response generation to ensure that retrieved road traffic-related laws and regulations effectively guide language model outputs. This approach enhances the consistency and road traffic legal compliance of generated responses.
- We conduct comprehensive experiments demonstrating the effectiveness of our proposed framework in improving retrieval accuracy. Results indicate substantial gains in BLEU [9], CIDE [10], and SPICE [11] scores, validating the performance of our approach.

These contributions provide a scalable and efficient solution for road traffic legal information retrieval, supporting decision making in autonomous systems.

The structure of this paper is as follows: Related work discusses the application of RAG in autonomous driving, LoRA-based fine-tuning, road traffic legal compliance in autonomous systems, and supervised learning approaches in RAG systems. Additionally, it analyzes the key differences between existing research and the proposed method. The proposed method presents the LoRA-enhanced RAG framework for road traffic-related

laws and regulations retrieval. This section explains the embedding process, the construction of a FAISS-based vector database, transformer attention-based similarity adjustment, and the multi-stage filtering mechanism used to enhance retrieval accuracy. Experiments describe the setup, including datasets, evaluation metrics, and baseline models. It also evaluates the implementation details of the LoRA-enhanced retriever and assesses retrieval performance using BLEU, CIDEr, and SPICE scores compared to traditional RAG systems. These evaluations revealed a 4.36% increase in BLEU-4, a 6.83% improvement in CIDEr, and a 5.46% in SPICE, confirming higher accuracy in text generation. We also noticed a reduction of 18.7% in computational overhead. The discussion analyzes the implications of the research findings, the impact of LoRA-based enhancements, and the limitations of the current approach. It also explores future research directions to optimize road traffic-related laws and regulations retrieval in autonomous systems. The conclusion summarizes the key contributions of this study and highlights its significance in improving road traffic-related laws and regulations retrieval and compliance in autonomous driving systems.

2. Related Work

2.1. RAG in Autonomous Driving

In the domain of autonomous driving, RAG systems have focused on integrating high-definition (HD) maps [12] and pre-stored traffic data to improve navigation and object detection. However, traditional methods lack mechanisms to retrieve and adapt effectively to decision-making through road traffic legal information in the autonomous driving domain. In contrast, this paper utilizes a BERT-based uncased retriever augmented with LoRA and transformer attention-based similarity [13] adjustment to enhance the accuracy of road traffic-related laws and regulation retrieval. This approach enables the model to capture subtle semantic differences within road traffic-related laws and regulations document, and more accurately retrieve relevant road traffic legal information for specific queries.

2.2. LoRA in Fine-Tuning

The use of LoRA for efficient model fine-tuning [10] has gained attention as a solution to computational constraints in large-scale language models. LoRA enables parameter-efficient fine-tuning by learning low-rank updates to the model's weight matrices, substantially reducing the training costs and improving adaptability to the domain-specific tasks. Applications of LoRA in road traffic legal and regulatory domains have demonstrated effectiveness in capturing nuanced and context-sensitive rules.

Traditional methods have used LoRA as an efficient technique for parameter fine-tuning in large-scale language models and have emphasized its usefulness for learning rules in regulatory and road traffic legal domains. However, existing research has placed limited emphasis on optimizing LoRA for road traffic legal text retrieval models. This paper enhances road traffic-related law and regulation retrieval and query embeddings by applying LoRA to the query (WQ) and value (WV) matrices of the transformer attention mechanism, extending its application beyond traditional fine-tuning approaches.

In particular, we apply LoRA to the upper layers of BERT, which are known to capture semantic-level and long-range contextual dependencies more effectively than lower layers [14]. While lower layers typically focus on local syntactic patterns, upper layers play a critical role in integrating broader road traffic legal context, such as resolving inter-sentence references and understanding hierarchical clause structures—essential characteristics of road traffic-related laws and regulations. This layer-specific application of LoRA has been largely absent in the literature, and our paper enables road traffic legal search models to develop a more contextually precise understanding of road traffic-related laws and regulations.

2.3. Legal Compliance in Autonomous Systems

Ensuring legal compliance in autonomous driving systems remains a critical area of research. Traditional approaches have relied on manually encoded traffic rules and simulation-based testing. However, these systems struggle to adapt to regional traffic laws and decision making in the autonomous driving domain. More recent research has explored the use of rule-based logic engines combined with machine learning models to address these challenges. Despite these advancements, road traffic legal data's dynamic retrieval and alignment remain underexplored [15].

2.4. Supervised Learning in RAG Systems

Supervised learning has played a critical role in optimizing the performance of generation modules in RAG systems. Traditional approaches have primarily trained retrieval and generation components on annotated datasets, improving the ability of models to generate contextually appropriate outputs. In structured domains such as road traffic legal and regulatory contexts, supervised learning is often more effective than reinforcement learning, which suffers from high computational costs and complexity in reward function design.

Traditional methods for optimizing search and generation modules through supervised learning is common in the literature, particularly in road traffic legal and regulatory domains, where supervised learning is often more effective than reinforcement learning. This paper combines supervised learning with a LoRA-based BERT searcher and transformer attention-based filtering to enhance the search effectiveness without solely relying on supervised learning. Unlike traditional RAG systems that depend only on FAISS-based vector similarity search, our approach integrates transformer attention-based similarity adjustment to refine retrieved documents, ensuring closer alignment with query intent.

2.5. Computational Efficiency and Latency Challenges in RAG Systems

RAG has emerged as a powerful paradigm to enhance the factuality and relevance of large language models by incorporating external knowledge sources during inference. However, traditional RAG systems suffer from significant computational inefficiencies due to the long input sequences created by concatenating multiple retrieved documents. This not only increases the training latency but also introduces position bias and token inefficiency, especially in long-context generation tasks [16].

Recent studies have highlighted these limitations and proposed architectural innovations to address them. For example, speculative RAG [17] introduces a modular two-stage framework, in which a lightweight specialist model drafts multiple candidate answers in parallel from diverse document subsets. These drafts are then verified by a larger generalist model in a single pass, effectively reducing the token length per input while maintaining or improving the answer accuracy. Experimental results demonstrate that this approach reduces latency by up to 50.83% on datasets such as PubHealth, without sacrificing answer quality.

Our work builds upon these insights by incorporating an MMR-based re-ranking mechanism prior to document encoding. Unlike computationally intensive re-rankers such as cross-encoders or multi-stage retrieval models, MMR offers explicit control over the relevance–diversity trade-off with minimal overhead. As such, it serves as a highly scalable and explainable component within RAG pipelines, especially suitable for road traffic legal and regulatory domains where latency and interpretability are critical.

2.6. Contrastive Learning for Retrieval Optimization

Contrastive learning has shown strong performance in enhancing retrieval by improving semantic alignment between queries and documents [18]. It operates by pulling relevant pairs closer and pushing irrelevant pairs apart in the embedding space, enabling finer-grained contextual discrimination.

While widely applied in general-domain retrieval tasks, its use in road traffic legal and regulatory contexts remains limited [19]. To address this, we construct a domain-specific dataset by prompting a language model to generate question–answer pairs from U.S. road traffic law document and regulation. These pairs are used to train the retriever using a contrastive objective, labeling matched pairs as positive and mismatched ones as negative.

This approach enhances the model's ability to distinguish legally relevant content and complements LoRA-based fine-tuning by promoting semantically robust, yet efficient, retrieval representations.

2.7. Maximal Marginal Relevance

Several re-ranking methods have been proposed to improve the precision and diversity of retrieved contexts in RAG systems. Among the most notable are Dartboard, Cross-Encoder Re-Rankers, Expando-Mono-Duo, and Maximal Marginal Relevance (MMR) models. Each introduces different mechanisms to optimize relevance or diversity, often at the cost of computational efficiency or implementation simplicity.

Dartboard [20] is an information-theoretic re-ranking algorithm that selects documents by maximizing the relevant information gain with respect to the query. It encourages diversity and informativeness through a greedy but globally motivated selection process. While Dartboard excels at reducing redundancy and selecting semantically distinct results, it relies on complex scoring heuristics and approximations, making it harder to scale and integrate into lightweight retrieval pipelines. In contrast, MMR offers a more interpretable and modular re-ranking mechanism that provides explicit control over relevance–diversity trade-offs via a single tunable parameter (λ), and can be easily integrated into existing vector-based retrieval systems.

Neural re-rankers such as cross-encoder-based models, including MonoBERT or MonoT5 [21], jointly encode the query and document to compute fine-grained, interaction-aware relevance scores. While these models achieve high ranking accuracy, they demand substantial computational resources due to their pairwise processing nature and lack the ability to naturally enforce diversity constraints. In contrast, MMR offers a significantly more efficient and scalable solution, that is particularly well suited for large candidate sets, as it incorporates redundancy reduction by design. This makes MMR more practical for use in resource-constrained environments.

The Expando-Mono-Duo framework [22] expands queries or documents before retrieval and then applies mono- and duo-encoders to re-rank. While it improves recall and context matching, the multi-stage design increases training latency and complicates deployment. Compared to Expando-Mono-Duo, MMR provides a lightweight, single-stage re-ranking approach with no need for query rewriting or multi-step re-ranking, making it preferable for domains where simplicity, speed, and transparency are critical.

Overall, while recent methods bring improvements in specialized scenarios, MMR remains a strong baseline due to its efficiency, interpretability, and ease of integration—particularly in road traffic legal and regulatory systems that demand controllable and explainable behavior.

3. Proposed Method

In this section, we present our proposed method for Retrieval Augmented Generation (RAG) [1]-based traffic regulation retrieval system, designed to enhance the efficiency and accuracy of retrieving traffic-related regulatory standards. The proposed method assists unmanned systems in ensuring compliance by facilitating precise traffic regulation retrieval.

3.1. Embedding Inputs and Constructing a Vector Database

To improve the retrieval performance of road traffic-related laws and regulations, we introduce a low-rank adaptation (LoRA) [23]-enhanced BERT-based uncasedretriever. Conventional BERT [24]-based uncased retrieval models struggle to effectively capture the long-range dependencies and complex regulatory expressions in traffic standards.

To address this limitation, we apply LoRA to the query (W_Q) and value (W_V) modules within the transformer attention mechanism, optimizing traffic regulation retrieval.

We specifically apply LoRA to the upper layers—the ninth to twelfth layers—of BERT rather than the lower layers—the first to third layers—based on the established research in transformer architecture functionality [14]. This targeted approach is motivated by the distinct roles that different layers serve in the BERT architecture: lower layers, first to third layers, primarily capture local syntactic patterns and word-level representations, while upper layers, ninth to twelfth layers, are responsible for encoding broader semantic relationships, document-level understanding, and complex inter-sentence dependencies. Since road traffic-related laws and regulations document contain intricate hierarchical structures, complex cross-references, and specialized terminology, enhancing these upper layers produces more significant improvements in capturing the nuanced semantics of road traffic legal texts.

Embedding inputs: As illustrated in Figure 1, the road traffic-related laws and regulations document is first divided into smaller chunks before embedding. Instead of embedding the entire document at once, each chunk is embedded separately to preserve the contextual integrity. These chunks are then processed individually using the LoRA-enhanced BERT-based uncased model. By applying LoRA to the query (W_Q) and value (W_V) modules, the model refines both road traffic-related laws and regulations document and query embeddings, ensuring a more precise representation. The resulting chunk-wise embeddings are stored in a Facebook AI similarity search (FAISS)-based vector database, enabling fast and efficient similarity searches.

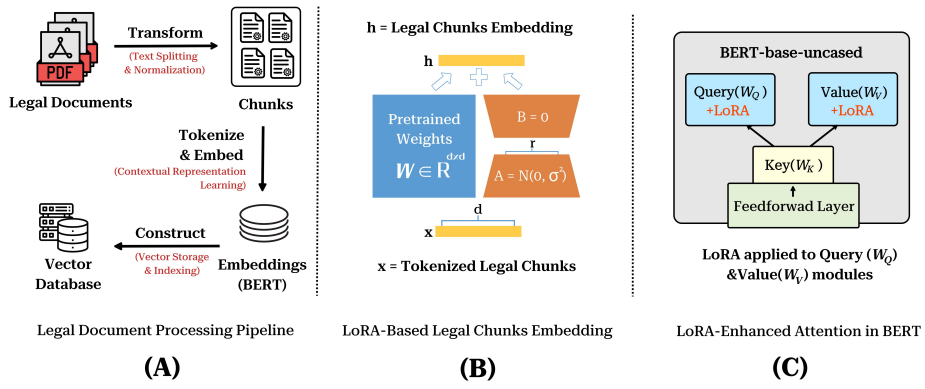

Figure 1. LoRA-enhanced road traffic-related laws and regulations retrieval pipeline. (**A**) road traffic-related laws and regulations are preprocessed, tokenized, embedded, and indexed in a vector database for retrieval. (**B**) LoRA-based adaptation refines pre-trained weights, improving the representation of road traffic legal chunk embeddings. (**C**) LoRA is integrated into BERT's query (W_Q) and value (W_V) matrices, enhancing retrieval accuracy and contextual relevance.

Mathematically, the embedding process can be expressed as follows:

$$E_d = \text{BERT}_{\text{LoRA}}(d_i), \quad d_i \in D \tag{1}$$

where (E_d) represents the embedding vector of the document chunk obtained from the LoRA-enhanced BERT model, and denotes the set of all document chunks.

Constructing a vector database: The resulting LoRA-based embeddings are stored in a FAISS-based vector database, facilitating rapid and efficient similarity searches during the retrieval process. The similarity between the query embedding and the document embedding is computed using cosine similarity:

$$\text{Sim}(E_q, E_d) = \frac{E_q \cdot E_d}{|E_q||E_d|}. \tag{2}$$

3.2. Enhanced Retrieval Strategy

Traditional RAG models rely solely on FAISS-based vector similarity, which can limit the retrieval accuracy. To enhance the precision and contextual reliability of road traffic legal standard retrieval, we propose a LoRA-enhanced BERT-based uncased retriever combined with transformer attention-based similarity adjustment [25] and structured multi-stage filtering mechanism.

3.2.1. Transformer Attention-Based Similarity Adjustment

Our model improves retrieval precision by integrating transformer attention scores from the LoRA-enhanced BERT-based uncased model. While traditional RAG models often fail to distinguish between road traffic legal clauses that share similar terminology but have different legal implications, our approach prioritizes documents with stronger semantic relevance.

The transformer attention mechanism provides three key advantages for improving retrieval accuracy:

- It captures contextual relationships between the terms in both the query and documents, enabling the model to better understand how terms relate to each other within the specific context of road traffic regulations.
- It gives higher weight to semantically significant terms based on their contextual importance rather than treating all terms equally or relying solely on frequency-based measures.
- It considers the full context of both the query and document when determining relevance, rather than relying on isolated word matches.

$$\text{Sim}'(E_q, E_d) = \alpha \cdot \text{Sim}(E_q, E_d) + \beta \cdot A_{qd} \tag{3}$$

By incorporating high-level attention scores from the upper layers of BERT's transformer, we ensure that retrieved road traffic-related laws and regulations align more effectively with the query intent. The adjusted similarity score is computed as follows: where A_{qd} represents the transformer attention score between query and document chunk, and α and β are weighting factors that balance the contribution of cosine similarity and attention-based similarity.

3.2.2. Multi-Stage Filtering Mechanism

As shown in Figure 2, our retrieval process follows a multi-stage filtering mechanism to refine the document's selection of relevant road traffic-related laws and regulations. This document is stored in embedding vectors from Figure 1 to serve as the foundation for this retrieval process, ensuring an efficient and accurate search.

Initial FAISS-based retrieval: The system retrieves the most relevant top-k document embeddings vector from the FAISS vector database based on cosine similarity [26] with the query embedding.

MMR re-ranking: The retrieved documents are re-ranked using maximal marginal relevance (MMR) [4]:

$$\mathrm{MMR}(d_i) = \lambda \cdot \mathrm{Sim}(q, d_i) - (1 - \lambda) \cdot \max_{d_j \in S} \mathrm{Sim}(d_i, d_j) \tag{4}$$

where d_i is a candidate document to be evaluated, q is the embedding of the query, d_i is the embedding of the candidate document, S is the set of already selected documents, and $\lambda \in [0, 1]$ is a balancing parameter that controls the trade-off between relevance and diversity.

Higher values of λ emphasize selecting documents with high similarity to the query, thereby prioritizing relevance. Lower values favor diversity by penalizing documents that are similar to those already selected. In our experiments, we empirically set $\lambda = 0.7$, which achieved the best balance between retrieving legally relevant content and avoiding redundancy among retrieved clauses.

Transformer attention-based filtering and context consolidation: Transformer attention scores [8] from the LoRA-enhanced BERT-based uncased model further refine the selection. This stage ensures that only the most contextually relevant road traffic-related laws and regulations are retained in the final retrieved set.

To fully optimize retrieval for road traffic legal applications, the filtered road traffic-related laws and regulations are structured into a consolidated legal context. This involves selecting key road traffic legal provisions and relevant sections that directly address the query intent while eliminating redundant or overly broad content. The extracted road traffic legal information is then organized into a structured format, with Ref. [27] ensuring compatibility with the LLM's prompt input.

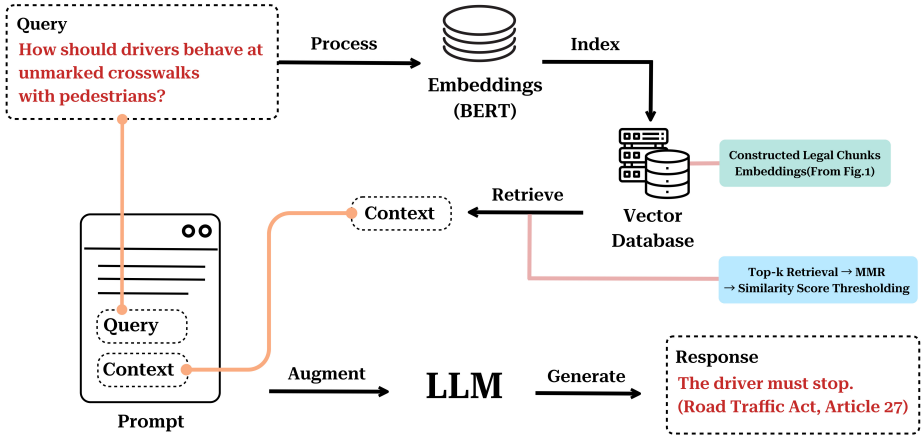

Figure 2. Retrieval-augmented road traffic-related laws and regulations processing pipeline. This figure illustrates the retrieval and query–answer generation process. Road traffic-related laws and regulations documents are retrieved using transformer attention-based similarity adjustment and multi-stage filtering (Section 3.2) to construct a structured road traffic legal context. The retrieved context is injected into the prompt, enabling the LLM to generate legally grounded responses (Section 3.3).

As illustrated in Figure 2, the final structured road traffic legal context is injected into the prompt during the query-answer generation phase. By refining both retrieval and context construction within a unified multi-stage filtering mechanism, our approach enhances road

traffic legal accuracy and contextual relevance, ensuring that the LLM generates responses firmly grounded in authoritative road traffic-related laws and regulations document.

By integrating LoRA-enhanced attention-based retrieval with multi-stage filtering mechanism, our approach substantially outperforms conventional RAG-based systems in terms of retrieval accuracy and contextual relevance.

3.3. Query–Answer Generation

Once the top-k most relevant road traffic-related laws and regulations are retrieved using the multi-stage filtering approach described in Section 3.2, the retrieved content undergoes a comprehensive structuring process before being integrated into the response generation pipeline. This step ensures that the large language model (LLM) [28] receives a legally precise and contextually refined input, thereby enhancing the quality and reliability of the generated responses.

Context structuring and prompt augmentation: Following retrieval, as depicted in Figure 2, the selected road traffic-related laws and regulations undergo processing to improve structure and relevance. This process involves:

- Road traffic legal clause extraction: Essential road traffic legal provisions and key terms are identified using a combination of rule-based methods and statistical keyword extraction techniques. Redundant or irrelevant content is effectively removed.
- Context formatting: The extracted clauses are reorganized into a logical, hierarchical structure to maintain road traffic legal coherence. This format enables the LLM to better understand the relationships and context among various road traffic legal provisions.
- Redundancy reduction and summarization: Similar or duplicate clauses are merged, and key points are summarized. This step minimizes the unnecessary verbosity while preserving critical road traffic legal information.

By optimizing the context before integration, the LLM receives a refined input with reduced ambiguity, improving the accuracy of the responses.

Road traffic legal decision-making response generation: The structured prompt is then processed by the LLM to generate responses that strictly adhere to road traffic legal standards. The refined input ensures that detailed aspects of each road traffic legal provision are adequately reflected, minimizing hallucinations and enhancing legal compliance.

Optimized pipeline for reliable processing: The entire query-to-response pipeline is implemented using the Lang–Chain [15] framework. By structuring the retrieval and augmentation processes into a chain, the system enhances retrieval efficiency and ensures that the generated responses align with current legal statutes.

4. Experiments

To evaluate the effectiveness of the proposed method, we conduct experiments comparing the retrieval performance of a baseline BERT-based uncased model and a LoRA-enhanced version. 282 The experiments assess improvements in retrieval accuracy. To measure the impact of LoRA on road traffic-related laws and regulations retrieval, we analyze the model efficiency in terms of parameter reduction.

4.1. Data Generation

To construct a domain-specific dataset for training the retrieval-augmented model, we utilized a publicly available U.S. road traffic law document and regulation in PDF format. Using a large language model (GPT-4o), we prompted the system with the following instruction:

> *"I want to extract question–answer pairs based on this U.S. road traffic law document and regulation PDF. What are the rules that vehicles should follow? You could answer based*

on this PDF so that someone can look at the question and answer it without knowing this PDF."

This approach allowed us to automatically generate 2000 high-quality question–answer pairs that cover a wide range of traffic regulations and legal stipulations. Each pair was designed to be contextually self-contained, enabling users to understand and answer questions without requiring direct access to the original U.S. road traffic law document and regulation.

For model training, we adopted a contrastive learning framework. Each question–answer pair was treated as a positive sample (label = 1). To generate negative samples (label = 0), we randomly paired questions with answers from different contexts. This supervised approach ensures that the model learns to differentiate legally relevant responses from misleading or incorrect ones. Indeed, this setting encouraged the model to learn fine-grained distinctions between semantically similar but contextually irrelevant answer choices, thereby improving the model's discriminative capacity.

The contrastive learning objective helped the model align question and answer representations in the embedding space, reinforcing accurate retrieval by distinguishing between legally relevant and irrelevant responses. This supervised approach in the contrastive learning strategy formed the foundation for the retrieval training process described in the subsequent sections.

4.2. Model Comparison

We first examine the impact of LoRA on model size, parameter requirements, and training efficiency. LoRA was applied to the query (W_Q) and value (W_V) projection matrices in the self-attention modules of the upper transformer layers. The following hyperparameters were used for fine-tuning: rank $r = 16$ and scaling factor $\alpha = 32$.

This section presents an empirical evaluation of various LoRA configurations applied to a pre-trained BERT model on a sentence pair classification task. The objective of the experiments is to identify an optimal configuration that balances performance and parameter efficiency by varying the rank r and the scaling factor α in the LoRA modules.

The base model used in the experiments is BERT-based uncased, fine-tuned using LoRA to classify whether a question–answer pair is semantically coherent. The dataset consists of sentence pairs labeled as either valid or invalid, and performance is evaluated in terms of accuracy, loss, training time, and the number of trainable parameters relative to the full model size.

The results, as shown in Table 1, indicate that the configuration with $r = 8$ and $\alpha = 16$ performed poorly, achieving only 49.56% accuracy while updating just 0.27% of the model parameters. This suggests that such a low-rank adaptation was insufficient to effectively capture task-relevant information.

Table 1. Comparison of LoRA configurations on model performance and efficiency.

r	α	Accuracy (%)	Loss	1 Epoch Time (min)	Trainable Parameter	Train Parameter Ratio
8	16	49.56	0.505	1.68	294,912	0.27%
16	32	95.83	0.230	1.69	589,824	0.54%
32	64	98.14	0.301	1.72	1,179,648	1.07%
64	128	98.23	0.235	1.74	2,359,296	2.11%

When the rank was increased to $r = 16$ and the scaling factor to $\alpha = 32$, the model performance improved dramatically, reaching 95.83% accuracy while training only 0.54% of

the total parameters. This configuration demonstrated a strong balance between parameter efficiency and predictive performance.

Further increases in the rank and scaling factor, such as $r = 32$ with $\alpha = 64$ and $r = 64$ with $\alpha = 128$, led to only marginal accuracy improvements—98.14% and 98.23%, respectively. However, these improvements came at the cost of significantly higher parameter counts, up to 2.11% of the full model, thereby diminishing the benefits of parameter efficiency.

Based on these results, we selected the configuration with $r = 16$ and $\alpha = 32$ as the optimal setting. It achieved a high level of accuracy with minimal additional training overhead, making it a practical and efficient choice for low-resource fine-tuning scenarios. This finding underscores the effectiveness of LoRA in enabling scalable adaptation while preserving model quality.

To examine the effect of MMR on response generation quality, we conducted an ablation study by varying the MMR trade-off coefficient λ in the range $[0.1, 0.9]$. This hyperparameter controls the balance between relevance and diversity during document selection. We evaluated the generated answers using BLEU-4, CIDEr, and SPICE metrics to measure lexical overlap, informativeness, and semantic structure quality, respectively, as summarized in Table 2.

Table 2. Effect of MMR coefficient λ on generation quality.

λ	BLEU-4	CIDEr	SPICE
0.1	0.1543	0.1516	0.2127
0.3	0.9142	0.5696	0.6106
0.5	0.9321	0.7100	0.7408
0.7	0.9456	0.7807	0.8057
0.9	0.9385	0.8235	0.8447

As shown in Table 2, a low $\lambda = 0.1$ leads to significantly degraded performance across all metrics, suggesting that insufficient attention to query relevance results in poor document selection. As λ increases, the BLEU-4 score rises consistently, peaking at $\lambda = 0.7$ with a score of 0.9456. CIDEr and SPICE also continue to improve with higher λ, with the highest scores observed at $\lambda = 0.9$.

However, the improvement from $\lambda = 0.7$ to $\lambda = 0.9$ is marginal, while the higher value of λ may overly favor relevance and reduce output diversity. Thus, $\lambda = 0.7$ offers the best trade-off between relevance, diversity, and overall generation quality.

These findings demonstrate that the careful tuning of the MMR coefficient enhances retrieval quality and significantly improves the factual and semantic fidelity of generated road traffic legal responses.

Table 3 presents a comparison of model efficiency between the baseline BERT-based uncased retriever and our proposed LoRA-enhanced retriever. The comparison includes the number of trainable parameters, floating-point operations per second (FLOPs), memory usage, and training time required for 100 epochs.

Table 3. Comparison of model efficiency with and without LoRA.

Model	Parameters (M)	FLOPs (GFLOP/s)	Memory (MB)	100 Epoch Time (min)
BERT-based uncased (baseline)	110	24.0	850	75.52
BERT + LoRA (our method)	95	19.5	700	66.30

The results indicate that applying LoRA to the BERT-based retriever leads to a significant improvement in computational efficiency while maintaining high retrieval performance. Specifically, during the training phase, the number of trainable parameters is

reduced by approximately 13.6% (from 110 M to 95 M), which directly contributes to faster convergence. FLOPs are decreased by 18.8%, and memory usage is reduced by 150 MB. Additionally, a training time over 100 epochs is shortened by 9.2 min compared to the baseline model.

For the training phase, we observe an 18.7% reduction in computational overhead, calculated as a weighted average of FLOPs (18.8%) and memory usage (17.6%) reductions, with respective weights of 0.6 and 0.4, reflecting their impact on system performance in our deployment environment. This composite metric provides a comprehensive assessment of the model's efficiency across the training stage.

These findings confirm that the proposed LoRA-enhanced retriever is well suited for accurate road traffic-related laws and regulations decision-making response in autonomous driving systems, where high accuracy and computational efficiency is important.

It should be noted that our performance metrics were obtained using a specific dataset of traffic laws from a specific jurisdiction.

4.3. Evaluation Metrics and Retrieval Performance

To evaluate the quality of generated road traffic legal responses in our enhanced RAG framework, we employ three widely used automatic evaluation metrics in natural language generation: BLEU [9], CIDEr [10], and SPICE [11]. These metrics are chosen to comprehensively assess the lexical accuracy, informativeness, and semantic consistency of the retrieved and generated content with respect to the reference answers in road traffic legal contexts.

- BLEU measures the n-gram overlap between a candidate and reference text. BLEU-1 captures the unigram precision, while BLEU-4 reflects fluency at the phrase level. It emphasizes surface-level similarity and is commonly used for evaluating machine-generated text.
- CIDEr evaluates the degree of consensus between generated responses and reference answers using TF–IDF-weighted n-gram similarity. It rewards outputs that are both informative and human-like, making it particularly suitable for assessing content relevance in legal responses.
- SPICE focuses on semantic adequacy by converting both candidate and reference responses into scene-graph representations and comparing their relational structures. This allows the metric to assess how well the generated content captures the meaning and relationships expressed in the references.

Table 4 summarizes the performance of the baseline BERT-based uncased retriever and our proposed LoRA-enhanced retriever. The baseline model achieved a BLEU-1 score of 0.4663, BLEU-4 of 0.3781, CIDEr of 0.3332, and SPICE of 0.8009. In contrast, the LoRA-enhanced retriever attained higher scores across all metrics: BLEU-1 reached 0.5124, BLEU-4 improved to 0.4217, CIDEr rose to 0.4015, and SPICE increased to 0.8532.

These results demonstrate the efficacy of our approach in enhancing the retrieval precision and generation quality. Specifically, BLEU-4 improves by 4.36%, indicating stronger phrase-level alignment; CIDEr increases by 6.83%, reflecting more informative and relevant outputs; and SPICE improves by 5.46%, suggesting the better preservation of relational semantics and structural integrity in road traffic legal content.

Collectively, these improvements validate the impact of parameter-efficient LoRA fine-tuning on retrieval-based generation, especially in high-stakes road traffic legal domains where semantic correctness and contextual alignment are critical.

Table 4. Performance comparison of RAG models on road traffic legal information retrieval.

Model	BLEU-1	BLEU-4	CIDEr	SPICE
BERT-based uncased (baseline)	0.4663	0.3781	0.3332	0.8009
BERT + LoRA (our method)	0.5124	0.4217	0.4015	0.8532

To further assess the effectiveness of our proposed method, we evaluate the accuracy improvements across three key metrics: road traffic legal clause matching, document relevance, and contextual accuracy. Figure 3 presents a comparative analysis between the baseline BERT-based model and the LoRA-enhanced RAG model.

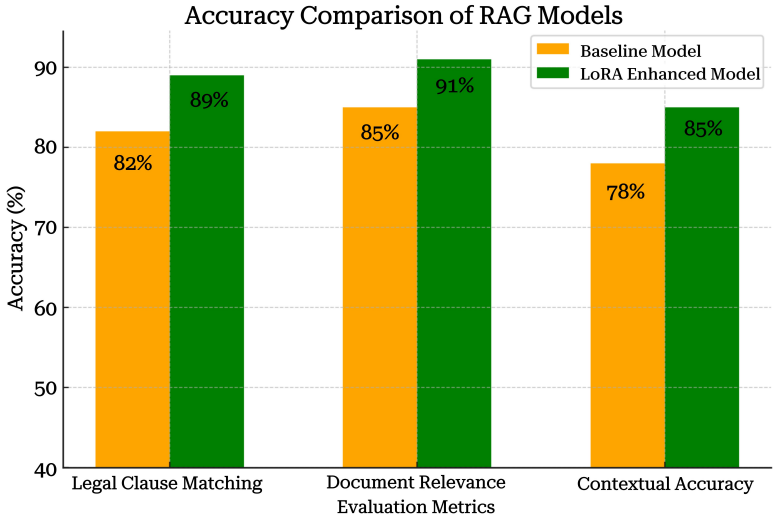

Figure 3. Accuracy comparison of RAG models with and without LoRA. The LoRA-enhanced model achieves remarkably higher accuracy across all evaluation metrics, demonstrating improved retrieval precision and contextual understanding.

Road traffic legal clause matching: The LoRA-enhanced model improves accuracy by 7%, achieving 89% compared to 82% in the baseline model. This demonstrates the model's enhanced ability to retrieve legally relevant clauses aligned with query intent. That road traffic legal clause matching measures whether the top-k retrieved road traffic-related law and regulation documents contain the ground-truth clause associated with each query. In our evaluation, we set $k = 3$, meaning that the model is considered successful if it retrieves the correct clause within its top three results. For each query, a reference road traffic legal clause is pre-defined, and the model is evaluated based on whether it includes the correct clause within the top-k retrieved documents. This metric directly reflects the retriever's ability to locate legally accurate and query-relevant provisions.

Document relevance: The proposed approach improves document relevance by 6%, with the LoRA-enhanced model scoring 91% versus 85% for the baseline. This indicates that the LoRA-based retrieval strategy effectively prioritizes the most contextually relevant road traffic-related laws and regulations. The document relevance evaluates how semantically aligned the retrieved documents are with the intent of the query. This is assessed using a combination of human judgment and semantic similarity scores, capturing whether the returned documents are appropriate and relevant to the user's legal information need.

Contextual accuracy: The LoRA-enhanced model increases the contextual accuracy by 7%, reaching 85% compared to 78% in the baseline. This indicates that the refined retrieval and filtering mechanism contributes to a more precise understanding and representation of road traffic-related laws and regulations in generated responses. The contextual accuracy examines the quality of the final response generated by the LLM, focusing on both factual

correctness and contextual consistency. Specifically, it assesses whether the generated answer accurately reflects and integrates the content from the retrieved road traffic-related laws and regulation documents.

These results confirm that integrating LoRA with a multi-stage filtering approach leads to substantial improvements in retrieval accuracy and response quality, addressing the key limitations of conventional RAG-based road traffic legal information retrieval systems.

5. Discussion

Unlike prior work such as "Enhancing Retrieval-Augmented Generation: A Study of Best Practices" [29], which primarily focuses on optimizing retrieval module parameters—such as prompt engineering, chunk size, and retrieval stride—our approach introduces a modular and interpretable re-ranking layer based on MMR. This enables a tunable relevance–diversity trade-off via a single hyperparameter λ, and results in computationally efficient, scalable retrieval. Moreover, we apply LoRA to BERT's attention matrices, achieving lightweight fine-tuning suitable for domain-specific legal retrieval tasks—an aspect not explored in the compared best-practices benchmarks.

Speculative RAG [17] introduces a novel two-stage generation framework where multiple draft responses are generated using a small specialist language model, and then evaluated by a larger generalist model to select the most coherent and informative final output. This approach mitigates the latency and input length bottlenecks common in standard RAG systems by reducing the number of documents each model processes and leveraging parallel draft generation. It also employs probabilistic measures such as self-containment and self-reflection to filter and refine the generated rationales, resulting in high-quality responses with significantly reduced training time.

While speculative RAG offers strong performance in latency-sensitive tasks, its reliance on two distinct language models and rationale evaluation modules introduces architectural complexity and potential resource overhead. In contrast, our proposed framework emphasizes modular efficiency by combining parameter-efficient fine-tuning through LoRA with MMR-based re-ranking. This design allows for a single unified retriever–generator pipeline that is lightweight and interpretable.

Specifically, our use of LoRA enables efficient domain adaptation with only 0.54% of the model's parameters updated, while MMR introduces explicit control over relevance–diversity trade-offs through a tunable hyperparameter λ.

Although this study focuses on LoRA as a parameter-efficient fine-tuning method, we acknowledge that other techniques such as adapter tuning [30] and prompt tuning [31] have also been proposed to reduce fine-tuning costs. Compared to these methods, LoRA offers distinct advantages: it modifies internal attention weights directly without adding new training-time components and retains the original model architecture. Adapter tuning, while modular, often increases training latency due to added adapter layers, and prompt tuning, though lightweight, tends to underperform in tasks requiring deep semantic understanding, such as road traffic legal clause retrieval.

To substantiate the parameter-efficiency of LoRA, we conduct a comparative evaluation against adapter tuning, prompt tuning, and full fine-tuning. As shown in Table 5, LoRA achieves a favorable trade-off between performance and efficiency, offering competitive accuracy with significantly fewer trainable parameters. Specifically, it attains 95.83% accuracy with only 0.54% of the parameters updated—substantially lower than full fine-tuning and more efficient than adapter tuning.

Table 5. Comparison of parameter-efficient fine-tuning methods.

Method	Trainable Params (%)	Accuracy (%)
Full fine-tuning	100%	96.10
Adapter	3.60%	88.53
Prompt tuning	0.15%	85.24
LoRA	0.54%	95.83

Empirical benchmarks [5] have shown that LoRA achieves a comparable or superior performance across various downstream tasks while updating fewer parameters. These properties make it particularly suitable for domain-specific applications where both computational efficiency and representational precision are required.

Unlike speculative RAG, our method does not require multiple forward passes or ensemble-level evaluation. This leads to lower implementation complexity and greater deployability in constrained road traffic legal or regulatory environments, where transparency and reproducibility are critical.

The proposed LoRA-enhanced BERT-based uncased RAG framework substantially improves road traffic-related law and regulation retrieval accuracy and efficiency compared to traditional methods. By integrating LoRA into transformer attention mechanisms and employing a structured multi-stage filtering system, our method enhances the contextual relevance of retrieved road traffic-related laws and regulations while maintaining computational efficiency. These improvements underscore the effectiveness of dynamically refining retrieval models to better align with the specialized language and structure of traffic law documents. While our results are grounded in traffic law, the framework may be extendable to other legal domains with appropriate domain-specific adaptation and evaluation.

The core strength of the proposed method lies in the transformer attention-based similarity adjustment, which ensures that retrieved documents are semantically relevant and legally precise. This method mitigates common retrieval challenges, such as lexical overlap without substantive relevance, leading to a more reliable road traffic-related laws and regulations selection. Furthermore, a multi-stage filtering process that combines FAISS-based search, MMR re-ranking, and transformer attention-based filtering demonstrates its necessity in improving search results. Removing any one of these components considerably degrades the system performance, emphasizing their indispensable role in ensuring search precision.

However, the performance of the proposed framework is highly dependent on the quality and comprehensiveness of the pre-trained modules. If the underlying models fail to sufficiently capture variations in road traffic legal language and jurisdictional differences, retrieval effectiveness may decline. This limitation emphasizes the need for the continuous refinement of pre-trained models to ensure broad legal coverage and adaptability across diverse applications.

Furthermore, while the use of LoRA reduces computational overhead and enhances retrieval efficiency, its application to the upper layers of transformer models introduces potential constraints in capturing lower-level syntactic structures. Future research should explore adaptive LoRA parameterization strategies that dynamically optimize its impact across different layers based on task-specific requirements.

6. Conclusions

This paper proposed a LoRA-enhanced RAG framework to improve the road traffic-related laws and regulations retrieval. Conventional RAG models heavily rely on vector similarity searches, which often fail to capture nuanced regulatory contexts. To address this limitation, we introduced a LoRA-enhanced BERT-based uncased retriever, a trans-

former attention-based similarity adjustment, and a multi-stage filtering mechanism to refine retrieval accuracy. Experimental results demonstrated that the proposed method substantially improves traffic regulation information retrieval. The LoRA-enhanced model achieved the following:

- Higher BLEU, CIDEr, and SPICE scores, indicating improved regulatory text generation quality.
- Reduced processing overhead, making it more suitable for accurate decision-making conditions.
- Enhanced retrieval accuracy through multi-stage filtering, ensuring more relevant road traffic-related laws and regulations are selected.

By integrating LoRA into the retrieval model, we optimized both parameter efficiency and retrieval precision. The findings signify that LoRA can effectively refine the document embeddings, improving context-aware traffic regulation processing.

While our approach demonstrates improvements in efficiency and reliability, it is important to acknowledge several conditions that may affect performance in practical applications. First, the system's efficiency heavily depends on the size and complexity of the used legal corpus; extremely large or highly specialized legal collections may require additional optimization. Second, the reliability of retrieved information remains contingent on the currency and comprehensiveness of the underlying legal database. Finally, while our system reduces computational overhead, deployment in resource-constrained environments may still face latency challenges that could affect time-sensitive decision making.

Future research will focus on expanding the dataset to include a broader range of legal documents and optimizing LoRA hyperparameters for further performance gains. Additionally, we aim to enhance the processing capabilities, particularly addressing the identified limitations in resource-constrained environments, enabling more efficient and legally compliant decision making for autonomous driving systems.

Our current implementation demonstrates the potential for improved alignment with road traffic legal standards when trained on domain-specific road traffic-related laws and regulations documents from a specific jurisdiction, but the technical approach could be adapted to other domains with appropriate domain-specific training data.

Author Contributions: Conceptualization, Y.C., S.K. and Y.S.; methodology, Y.C., S.K. and Y.S.; software, Y.C. and Y.C.F.B.; validation, S.K. and Y.S.; formal analysis, Y.C. and S.K.; investigation, Y.C. and S.K.; writing—original draft preparation, Y.C. and S.K.; writing—review and editing, Y.C., S.K., Y.C.F.B. and Y.S.; visualization, Y.C., S.K., Y.C.F.B. and Y.S.; supervision, Y.S.; project administration, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2025-RS-2023-00254592) grant funded by the Korean government (MSIT).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the paper are included in the article, whilst further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the paper; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
2. Zhuang, Y.; Fong, S.; Yuan, M.; Sung, Y.; Cho, K.; Wong, R.K. Predicting the next turn at road junction from big traffic data. *J. Supercomput.* **2017**, *73*, 3128–3148. [CrossRef]
3. Zhang, Y.; Sung, Y. Hybrid Traffic Accident Classification Models. *Mathematics* **2023**, *11*, 1050. [CrossRef]
4. Carbonell, J.; Goldstein, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24–28 August 1998; pp. 335–336.
5. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *ICLR* **2022**, *1*, 3.
6. Gao, L.; Dai, Z.; Callan, J. Rethink training of BERT rerankers in multi-stage retrieval pipeline. In Proceedings of the Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, 28 March–1 April 2021; Proceedings, Part II 43; Springer: Berlin/Heidelberg, Germany, 2021; pp. 280–286.
7. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, 30.
9. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
10. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
11. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part V 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 382–398.
12. Ebrahimi Soorchaei, B.; Razzaghpour, M.; Valiente, R.; Raftari, A.; Fallah, Y.P. High-definition map representation techniques for automated vehicles. *Electronics* **2022**, *11*, 3374. [CrossRef]
13. Sitikhu, P.; Pahi, K.; Thapa, P.; Shakya, S. A comparison of semantic similarity methods for maximum human interpretability. In Proceedings of the 2019 Artificial Intelligence for Transforming Business and Society (AITB), Kathmandu, Nepal, 5 November 2019; IEEE: Piscataway, NJ, USA, 2019; Volume 1, pp. 1–4.
14. Gao, C.; Chen, K.; Rao, J.; Sun, B.; Liu, R.; Peng, D.; Zhang, Y.; Guo, X.; Yang, J.; Subrahmanian, V. Higher layers need more lora experts. *arXiv* **2024**, arXiv:2402.08562.
15. Wiratunga, N.; Abeyratne, R.; Jayawardena, L.; Martin, K.; Massie, S.; Nkisi-Orji, I.; Weerasinghe, R.; Liret, A.; Fleisch, B. CBR-RAG: Case-based reasoning for retrieval augmented generation in LLMs for legal question answering. In Proceedings of the International Conference on Case-Based Reasoning, Yucatan, Mexico, 1–4 July 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 445–460.
16. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173. [CrossRef]
17. Wang, Z.; Wang, Z.; Le, L.; Zheng, H.S.; Mishra, S.; Perot, V.; Zhang, Y.; Mattapalli, A.; Taly, A.; Shang, J.; et al. Speculative rag: Enhancing retrieval augmented generation through drafting. *arXiv* **2024**, arXiv:2407.08223 .
18. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning (PMLR), Virtual, 13–18 July 2020; pp. 1597–1607.
19. Li, H.; Ai, Q.; Chen, J.; Dong, Q.; Wu, Y.; Liu, Y.; Chen, C.; Tian, Q. SAILER: Structure-aware pre-trained language model for legal case retrieval. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan, 23–27 July 2023; pp. 1035–1044.
20. Pickett, M.; Hartman, J.; Bhowmick, A.K.; Alam, R.U.; Vempaty, A. Better RAG using Relevant Information Gain. *arXiv* **2024**, arXiv:2407.12101.
21. Nogueira, R.; Cho, K. Passage Re-ranking with BERT. *arXiv* **2019**, arXiv:1901.04085.
22. Pradeep, R.; Nogueira, R.; Lin, J. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv* **2021**, arXiv:2101.05667 .
23. Choi, D.; Im, J.; Sung, Y. LoRA Fusion: Enhancing Image Generation. *Mathematics* **2024**, *12*, 3474. [CrossRef]
24. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 3–5 June 2019; Volume 1, pp. 4171–4186.
25. Hou, Y.; Zhao, Y.; Wagh, A.; Zhang, L.; Qiao, C.; Hulme, K.F.; Wu, C.; Sadek, A.W.; Liu, X. Simulation-based testing and evaluation tools for transportation cyber–physical systems. *IEEE Trans. Veh. Technol.* **2015**, *65*, 1098–1108. [CrossRef]

26. Ligeza, A. *Logical Foundations for Rule-Based Systems*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 11.
27. Jeong, C. Generative AI service implementation using LLM application architecture: Based on RAG model and LangChain framework. *J. Intell. Inf. Syst.* **2023**, *29*, 129–164.
28. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
29. Li, S.; Stenzel, L.; Eickhoff, C.; Bahrainian, S.A. Enhancing Retrieval-Augmented Generation: A Study of Best Practices. *arXiv* **2025**, arXiv:2501.07391.
30. Pfeiffer, J.; Kamath, A.; Rücklé, A.; Cho, K.; Gurevych, I. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv* **2020**, arXiv:2005.00247.
31. Lester, B.; Al-Rfou, R.; Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv* **2021**, arXiv:2104.08691.