RAG based Question Answering of Korean Laws and Precedents

Kiho Seo
Central Police Academy
Korean National Police Agency
seokiho_@_police.go.kr

Takehito Utsuro
Dept. of Intelligent Interaction Tech.,
Faculty of Engineering, Information and
Systems, University of Tsukuba
utsuro_@_iit.tsukuba.ac.jp

Abstract

We propose a method of improving the performance of question answering based on the interpretation of criminal law regulations in the Korean language by using large language models. In this study, we develop a system that accumulates legislative texts and case precedents related to criminal procedures published on the Internet. The system searches for relevant legal provisions and precedents related to the query under the RAG (Retrieval-Augmented Generation) framework. It generates accurate responses to questions by conducting reasoning through large language models based on these relevant laws and precedents. As an application example of this system, it can be utilized to support decision making in investigations and legal interpretation scenarios within the field of Korean criminal law.

1 Introduction

In recent years, the utilization of Large Language Models (LLMs) (Singh, 2024) in the legal field has been attracting attention, and their potential is particularly expected in legal interpretation and case analysis. In the Korean criminal justice system, as legislative amendments and the accumulation of precedents have led to a continuous increase in legal information that should be referenced, it is not easy for investigators to make prompt and accurate decisions. Here, conventional keyword-based search systems cannot sufficiently consider the context and semantic relevance of legal documents, making it difficult to efficiently acquire knowledge, particularly when educating and training new investigators on the effective use of keyword-based search systems.

Based on this background, this paper aims to model question answering based on criminal procedure-related legal interpretations and case references by utilizing the latest LLM technology. Specifically, we apply the RAG (Retrieval-Augmented Generation) (Lewis et al., 2021) framework to legal provisions and case information published on the Korean legal information website[1]. In the RAG framework, legal and case information is converted into embedding vectors and stored in a searchable database using FAISS[2]. Then, for a given question, the system searches for legal provisions and precedents related to the question, and based on these relevant legal provisions and precedents, generates accurate answers to questions by conducting reasoning through large language models. Through the RAG framework, it is expected to mitigate the hallucination problems of LLMs while improving the quality and efficiency of decision-making in investigations.

However, deploying RAG systems in legal domains carries significant risks, particularly hallucination — generating plausible but incorrect legal advice. Recent incidents have demonstrated that legal AI systems can provide misleading advice to users. Our comprehensive error analysis addresses these concerns by systematically categorizing failure modes in legal response generation, evaluating the reliability of RAG systems when generating applicable statutes, relevant precedents, and legal opinions through prompt engineering.

This paper makes the following key contributions:

- **Hierarchical Document Segmentation**: We develop a comprehensive three-tier hierarchical document segmentation methodology (articles, paragraphs, items) specifically designed for Korean legal texts, enabling fine-grained retrieval and legal reasoning.
- **Domain-Specific Query Expansion**: We propose a custom query expansion technique tailored to Korean legal terminology and multi-

[1] http://www.law.go.kr
[2] https://faiss.ai

Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER), pages 91–100
July 31, 2025 ©2025 Association for Computational Linguistics

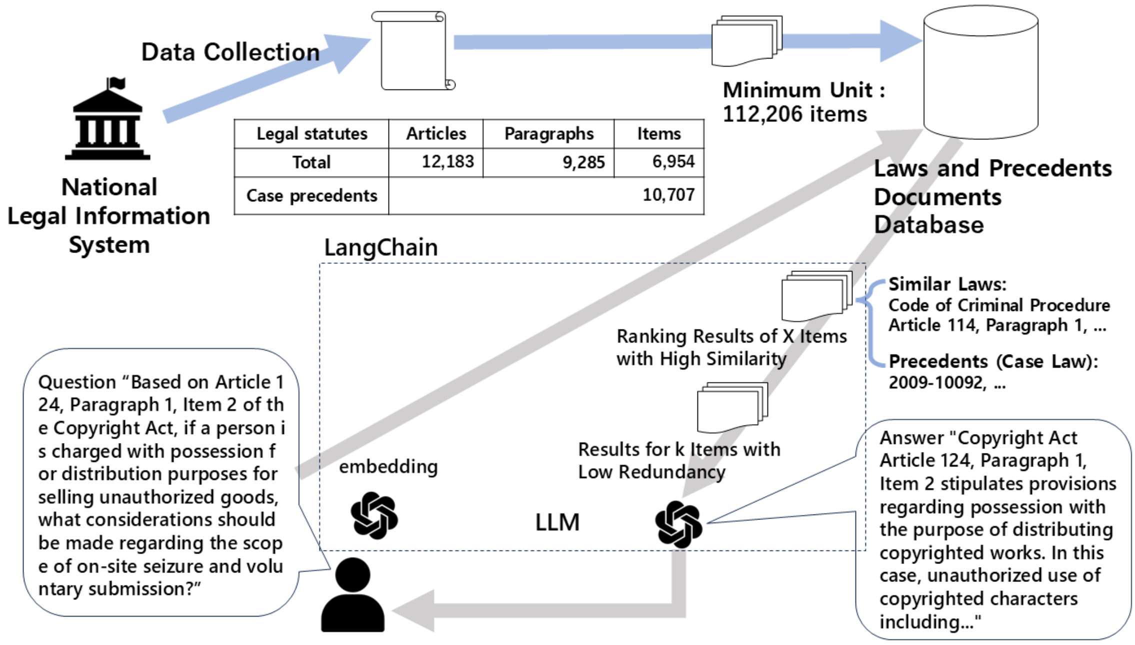

Figure 1: Architecture of the Korean Legal RAG System. The framework integrates (1) data collection from national legal databases, (2) hierarchical document segmentation and embedding, (3) query expansion and MMR-based retrieval, and (4) GPT-4-turbo based answer generation with legal reasoning.

statute queries, allowing more effective retrieval of relevant statutes and precedents for complex legal questions.

- **Systematic Error Analysis**: We conduct the first comprehensive error analysis of RAG systems in the Korean legal domain, identifying precedent selection as the primary challenge (23.7% of errors) while confirming high reliability in legislation application (0.8% error rate).

2 Related Work

In relation to this paper, literature (Hendrycks et al., 2021) publishes a dataset annotated by experts for the task of extracting important sections in legal contracts, and evaluates the performance of various models in the task. Literature (Papaloukas et al., 2021) proposed a model for topic classification of legal texts at various granularities of legal topics. Literature (Niklaus et al., 2021) proposes a multilingual legal judgment prediction benchmark using case precedent data. Literature (Hong et al., 2021) conducts research on information extraction tasks targeting dialogue examples in the legal domain. The literature (Choi et al., 2023) analyzes AI use methods in legal consultations and document preparation by lawyers. Literature (Trozze et al., 2024) describes the results of evaluating the usefulness of large language models as tools for legal interpretation and lawyer support in litigation.

3 Legal Statutes and Judicial Precedent Data

In Korea, the National Legal Information System[3] provides metadata for legal provisions and administrative regulations. In this paper, we use the API of this system to collect and utilize information on legal provisions (articles, paragraphs, items) and case precedents. As a result, as shown in Figure 1 and Table 1, we collected a total of 12,183 articles, 9,285 paragraphs, 6,954 items, and information on 10,707 criminal case precedents from 1975 to 2023. Next, regarding these provisions and precedents, we divided them by articles, paragraphs, and items, and saved them in the searchable database at the smallest unit, resulting in 112,206 minimum items being stored. We collected data from 20 core legal statutes related to criminal procedures. However, we identified 2,419 precedents classified as "Others" that either applied multiple statutes simultaneously or referenced statutes not included in our core 20 due to legal amendments or name changes. These additional legal statutes were included in our database to ensure comprehensive coverage, bringing the total number of statutory provisions

[3] https://open.law.go.kr

to 12,183 articles as shown in Table 1. Also, newly enacted regulations such as "Regulations on Mutual Cooperation between Prosecutors and Judicial Police Officers and General Investigation Rules" did not have corresponding precedents.

3.1 Legal Statutes

Our legal statute collection follows a hierarchical three-tier structure: articles, paragraphs, and items. Each tier serves as an independent searchable unit while maintaining hierarchical relationships.

Article Level: The highest structural unit containing complete legal provisions. For example, Criminal Act Article 43 (Sentence and Loss or Suspension of Qualification) constitutes a single article object with metadata including: title ("Sentence and Loss or Suspension of Qualification"), law name ("Criminal Act"), article number ("43"), enforcement date, and unique identifier ("Criminal_Act_43").

Paragraph Level: Subdivisions within articles marked by circled numbers (①, ②, etc.). For instance, "① A person who has been sentenced to death penalty, imprisonment for life, or imprisonment without prison labor for life shall lose the following qualifications:" forms a paragraph object with law name, article number, paragraph identifier (①), and unique identifier ("Criminal_Act_43_①").

Item Level: The most granular subdivisions within paragraphs, marked by numbers (1, 2, etc.). For example, "1. Qualification to become a public official" creates an item object with item number ("1") and unique identifier ("Criminal_Act_43_①_1").

Hierarchical Relationships: Articles can exist independently, but paragraphs require parent articles. Items can exist under articles with or without intermediate paragraphs. This structure enables both broad contextual searches and precise legal provision retrieval.

3.2 Judicial Precedent Data

The data structure for cases integrates case content and metadata in a unified format. For example, it is structured in formats such as "Case Number: 2023 No. 2102," "Case Name: Violation of the Act on the Control of Narcotics (Psychotropic Substances)," "Summary of Judgment: The description of the facts in the prosecution specifies the facts by clearly indicating the time, place, and method of the crime (omitted)." This structure includes information such as case number, case name, court name, judgment date, and referenced legal provisions, which are centrally managed as metadata. Additionally, the case content includes the case number, referenced legal provisions, and detailed descriptions of the precedent, enabling efficient searching and utilization of case information.

4 Question Answering System Using RAG

In this paper, for prompts in RAG, we use zero-shot and few-shot approaches (with evaluations conducted using 5-shot in this paper), and for embeddings in LLM, we use `text-embedding-ada-002`[4]. For the LLM itself, we use the GPT model `gpt-4-turbo`[5]. Additionally, we implement RAG using LangChain[6] as the platform.

4.1 Self-query Approach

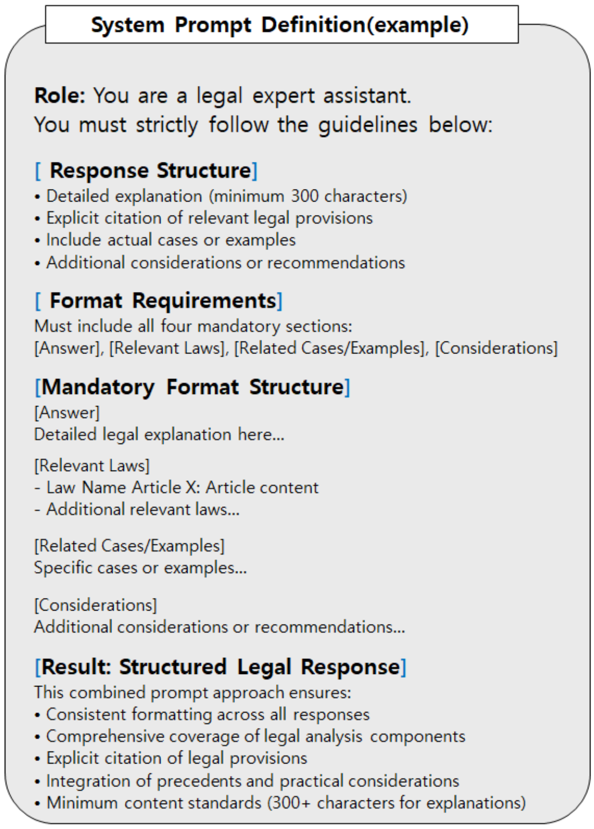

Figure 2: Legal Expert Assistant Prompt System Architecture

[4]https://platform.openai.com/docs/guides/embeddings#embedding-models

[5]https://platform.openai.com/docs/models#gpt-4-turbo-and-gpt-4

[6]https://www.langchain.com/

Legal Statute Name	#Art.	#Para.	#Items	#Cases	Total
Criminal Act	459	312	18	5,276	
Criminal Procedure Act	641	866	150	2,059	
Regulations on Mutual Cooperation between Prosecutors and Judicial Police Officers	87	174	98	0	
Special Act on Telecommunications Financial Fraud Prevention	30	69	83	3	
National Sports Promotion Act	107	251	156	9	
Korea Racing Authority Act	86	121	122	2	
Special Act on Prevention of Insurance Fraud	19	18	6	1	
Act on Prohibition of Improper Solicitation and Receipt	31	72	82	19	
Road Traffic Act	223	474	427	274	
Act on Special Cases for Traffic Accidents	6	8	15	45	
Dishonored Checks Control Act	7	8	3	69	
Attorney-at-Law Act	189	305	150	88	10,283
Specialized Credit Finance Business Act	129	267	242	7	
Special Measures Act on Real Estate Registration	12	21	10	4	
Act on Information Network Utilization and Protection	155	298	326	17	
Act on Punishment of Sexual Violence Crimes	68	192	85	165	
Act on Aggravated Punishment of Economic Crimes	14	35	12	10	
Act on the Punishment of Violence	10	21	14	215	
Act on Protection of Children Against Sexual Abuse	92	226	203	20	
Act on Punishment of Child Abuse Crimes	78	182	85	5	
Others	9,740	5,365	4,667	2,419	
Total	12,183	9,285	6,954	10,707	10,283

Table 1: Number of Articles, Paragraphs, Items, Cases, and Total Incidents by Legal Statute

Method	Response Type	Accuracy
5-shot, with Self Query	Ambiguous	90
	Unambiguous	137
	Total	60.4% (137/227)
5-shot, without Self Query	Ambiguous	87
	Unambiguous	140
	Total	**61.7**% (140/227)
zero-shot, with Self Query	Ambiguous	102
	Unambiguous	125
	Total	55.1% (125/227)
zero-shot, without Self Query	Ambiguous	100
	Unambiguous	127
	Total	55.9% (127/227)

Table 2: Evaluation Results

Query expansion enhances document retrieval by transforming a single query into multiple similar variants in LLMs. This technique generates alternative perspectives while preserving the original intent, allowing RAG models to leverage questions and context as cues for optimal query formulation, thereby improving information retrieval quality.

We developed a custom prompt engineering methodology for query expansion that generates multiple related queries from a single input question. Unlike LangChain's built-in Self Querying[7] API, our approach uses domain-specific prompt templates that explicitly instruct the model to generate follow-up questions addressing key legal elements such as constituent elements of crimes and culpability assessment.

Implementation Details: The system employs a two-stage approach for comprehensive legal analysis. First, the self-querying mechanism processes the original query and generates 3–5 additional related queries using domain-specific prompt templates that address key legal analysis components such as constituent elements, culpability assessment, procedural requirements, and precedent applicability. All generates queries are then used simultaneously for document retrieval, enabling comprehensive coverage of complex legal scenarios that span multiple statutes. Following the retrieval process, the system applies a structured prompt template as shown in Figure 2 to ensure consistent and comprehensive responses. The prompt defines the assistant as a legal expert that must strictly follow specific instructions, requiring responses to include four mandatory sections: [Answer] providing detailed legal explanation(minimum 300 characters), [Relevant Laws] with explicit citation of legal provisions, [Related Cases/Examples] describing specific precedents and examples, and [Considerations] offering additional recommendations. This structured approach ensures that all generated responses maintain consistent formatting and comprehensive coverage of legal analysis components.

Specific Example: When presented with the query "If a person is charged with possession

[7] https://python.langchain.com/docs/how_to/self_query/

for distribution purposes for selling unauthorized goods based on Article 124, Paragraph 1, Item 2 of the Copyright Act, what are the considerations for on-site seizure scope and voluntary submission?", our system generates complementary queries including: "What constitutes possession for distribution under copyright law?", "What are the procedural requirements for on-site seizure in intellectual property cases?", and "What precedents exist for voluntary submission in copyright violation cases?"

4.2 Extraction of Similar Documents Using MMR

Furthermore, in this paper, to extract documents relevant to questions, we adopt a method to search for multiple highly relevant documents by using MMR (Maximal Marginal Relevance) (Carbonell and Goldstein, 1998)[8] in LangChain. The formulation of MMR is shown below.

$$\mathrm{MMR} = \underset{D_i \in D \setminus S}{\arg\max} \Big[\lambda \cdot \mathrm{Sim}(D_i\text{fi}Q) - (1-\lambda) \cdot \max_{D_j \in S} \mathrm{Sim}(D_i\text{fi}D_j) \Big]$$

- λ: Weight between similarity and diversity($0 \leq \lambda \leq 1$)
- D: Set of all candidate documents for search
- S: Set of already searched documents
- $\mathrm{Sim}(D_i\text{fi}Q)$: Similarity between document D_i and Q
- $\mathrm{Sim}(D_i\text{fi}D_j)$: Similarity between documents D_i and D_j

This method allows for the selection of documents containing more diverse information while avoiding redundancy. For the similarity measure, we use the sum of SBERT[9] and BERTScore[10].

Parameter Optimization: For the main parameters λ (in increments of 0.1), X (number of candidate items to remove redundant results in search results), and k (number of results after removing redundant results in search results), we performed parameter adjustment and evaluation through two-fold cross-validation on 227 evaluation question-answering cases. Our experimental results demonstrated that the optimal performance on the experimental dataset was achieved with parameter values of $X = 10$, $k = 5$, and $\lambda = 0.9$ across zero-shot, few-shot, and self-query approaches.

[8] https://python.langchain.com/docs/how_to/example_selectors_mmr/

[9] https://sbert.net

[10] https://pypi.org/project/bert-score

5 Evaluation

5.1 Overview

For 227 evaluation question-answering cases, the first author manually evaluated the answers generated by the proposed method, classified whether the answers were ambiguous or clear, and then determined whether the answers were correct. The calculated accuracy results are shown in Table 2.

The assistance system utilizes LangChain's MMR retriever to extract five non-redundant case precedents per query. Evaluation against content relevance, legislation application, and precedent references revealed that 51.9% of errors occurred across all three criteria simultaneously. Precedent selection emerged as the primary challenge, with errors involved incorrect precedents only, while legislation application proved most reliable (only 0.8% of errors). Few-shot prompting without self-query achieved the highest accuracy, consistently outperforming Zero-shot approaches, though self-query implementation showed no significant improvement. Testing with 227 real-world cases confirmed performance improvements.

While our approach shows promise compared to traditional keyword-based retrieval methods, this study did not include baseline comparisons with conventional legal search systems. Future work will focus on systematic baseline comparisons to quantify performance improvements, expanding the dataset, developing systems for legislative and precedent updates, and refining precedent selection algorithms.

5.2 Error Analysis

Response accuracy was evaluated against three criteria: content relevance, correctly applied legislation, and appropriately referenced precedents. The analysis revealed that 51.9% of erroneous responses exhibited failures across all three criteria simultaneously, as shown in Figure 3, representing the most significant error category.

Among the remaining errors, those involving incorrectly referenced precedents alone constituted 23.7% of cases as shown in Figure 5. These errors were particularly notable for citing nonexistent case numbers or irrelevant judicial precedents despite providing correct legislative applications. This error pattern was observed in public indecency cases where defendants engaged in masturbation in semi-public spaces such as apartment balconies and hallways. The system correctly identified relevant

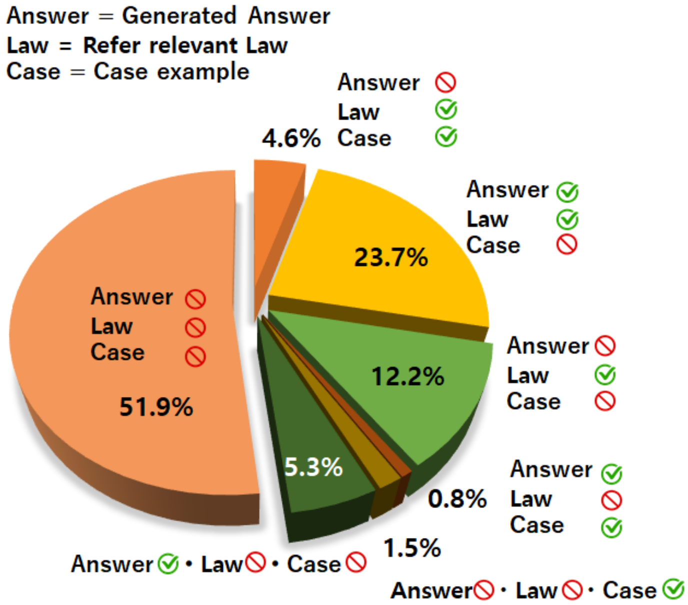

Figure 3: Error Distribution Analysis. Multi-criteria failures (51.9%) represent the largest error category, where responses fail across content relevance, legislation application, and precedent reference simultaneously. Precedent-only errors (23.7%) constitute the second largest category, highlighting precedent selection as a primary challenge.

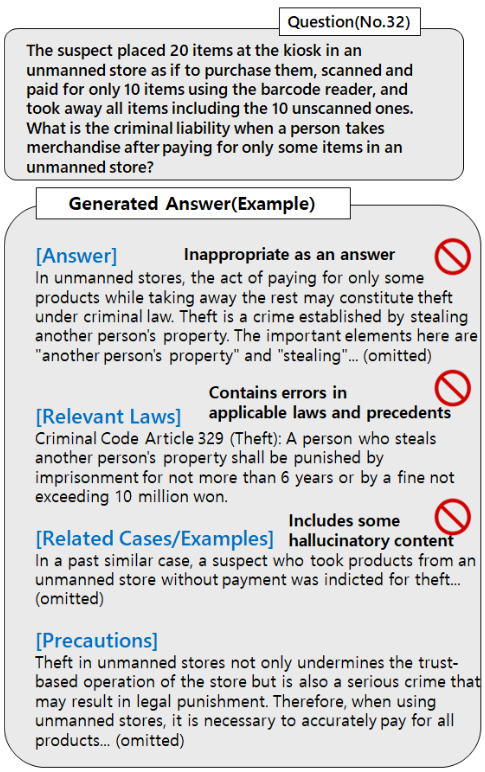

Figure 4: Error category = incorrect answer / irrelevant law / unrelated case-example

statutes but failed to cite appropriate precedential authority, instead referencing non-existent or unrelated case law.

Answer errors combined with precedent misidentification accounted for 12.2% of erroneous responses as shown in Figure 6. In these instances, while the applied legislation was correctly identified, the response content failed to appropriately address the query, and the cited precedents were irrelevant to the question posed. This error pattern was observed in domestic cohabitation disputes involving refusal to vacate shared residences, where the system struggled to provide coherent legal analysis despite correct statutory identification. This error pattern typically occurred in cases involving domestic disputes and housing law.

Less frequent error patterns included cases where content was correct but both legislation and precedents were incorrectly identified (5.3%) as shown in Figure 7. These responses demonstrated accurate understanding of the legal question but failed to properly ground the analysis in relevant legal frameworks. Such errors were observed in cases involving insurance liability and vehicle operation regulations.

Instances where only the response content was inaccurate (4.6%) as shown in Figure 8, represented scenarios where both legislation and precedent citations were correct, but the analysis contained sub-

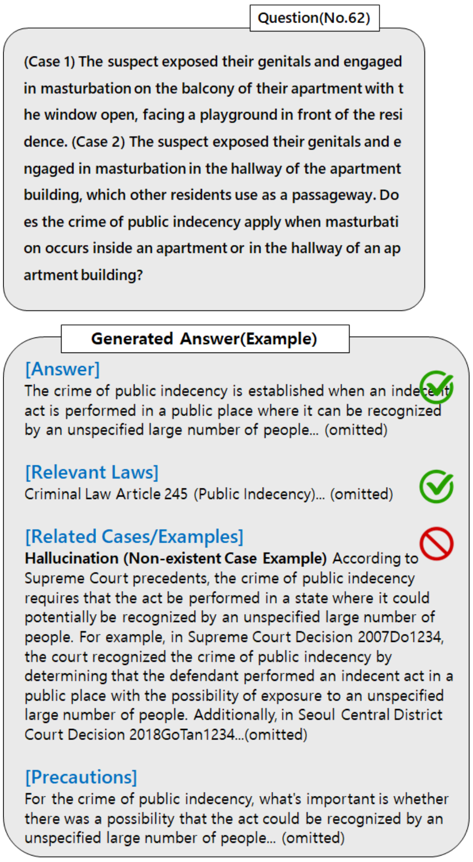

Figure 5: Error category = unrelated case-example only

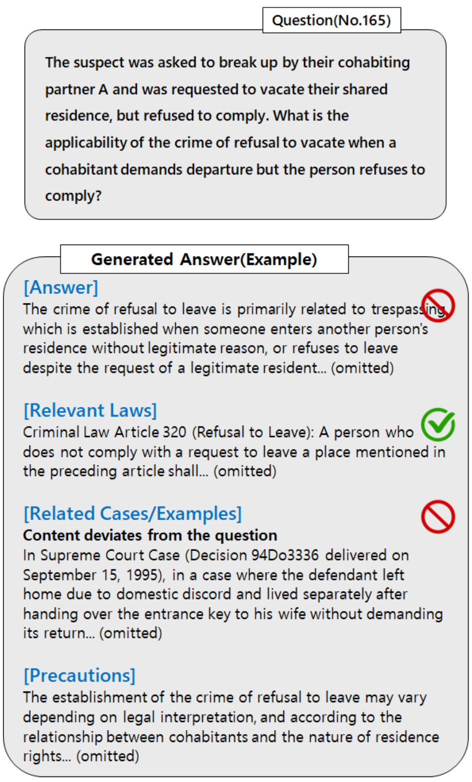

Figure 6: Error category = relevant law only

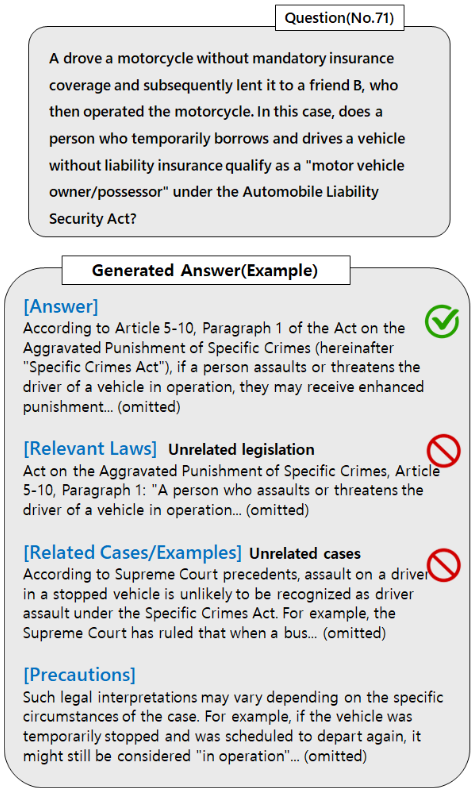

Figure 7: Error category = correct answer only

stantive errors. This pattern was observed in theft cases involving partial payment at unmanned stores, where the system correctly identified relevant laws and precedents but provided flawed legal reasoning. This pattern appeared in cases involving commercial transactions and theft-related offenses.

Scenarios where only precedent selection was accurate (1.5%) as shown in Figure 9,involved responses where both the content and applied legislation were incorrect, yet the system successfully identified relevant case precedents. This pattern was observed in healthcare fraud cases involving identity deception for medical treatment and insurance benefits, where appropriate precedents were cited despite incorrect statutory analysis and flawed reasoning. This rare pattern was observed in cases involving identity fraud and healthcare-related offenses.

The rarest category involved cases where legislation application alone was incorrect (0.8%) as shown in Figure 10. In these instances, both the response content and referenced precedents were appropriate, but the system erroneously cited inappropriate statutory provisions. This pattern occurred in drunk driving cases involving breath test refusal due to physical injury, where the system

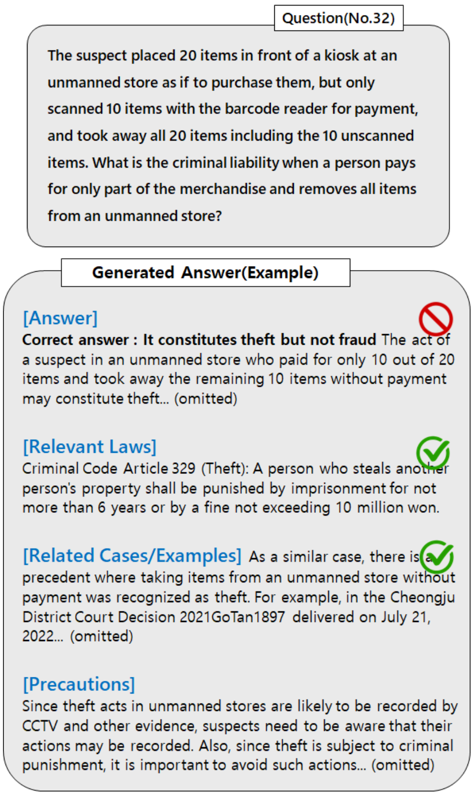

Figure 8: Error category = incorrect answer only

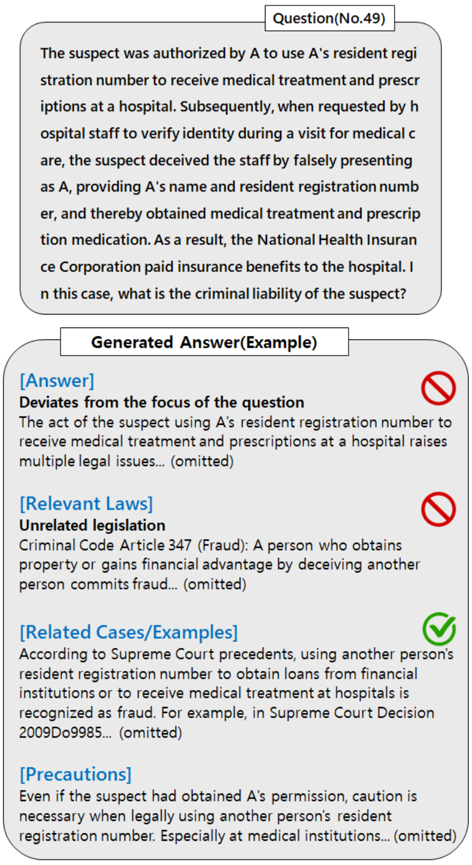

Figure 9: Error category = related case-example only

provided sound analysis and relevant precedents but incorrectly applied statutory measurement refusal provisions. This pattern appeared in cases involving traffic violations and alcohol-related offenses.

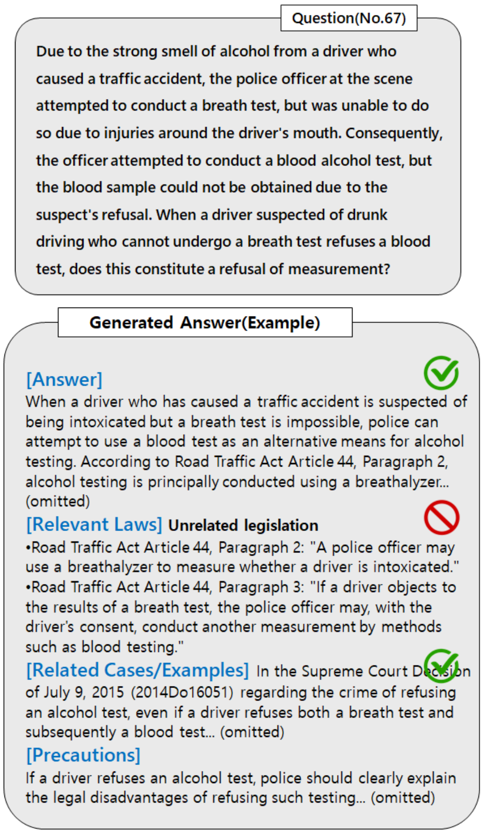

Figure 10: Error category = irrelevant law only

This comprehensive error analysis demonstrates that precedent selection represents the primary challenge affecting system accuracy, while legislation application demonstrates the highest reliability when evaluated as an independent component. The significant proportion of multi-criteria failures suggests potential interdependencies in the system's reasoning process that warrant further investigation. As shown in the error analysis figures (Figures 2-8), the system's performance varies significantly depending on the type of legal question and the complexity of the required legal reasoning, with precedent selection being particularly challenging in cases that involve recent or nuanced legal interpretations.

6 Conclusion

This study demonstrated the effectiveness of a retrieval-augmented generation approach for Korean criminal law question answering. By integrating legislative texts and judicial precedents, the proposed framework enables context-aware legal reasoning. In evaluations on 227 real-world cases, few-shot prompting consistently outperformed zero-shot prompting, achieving an accuracy of 61.7%. Error analysis indicated that precedent selection was the primary source of errors (23.7%), while legislation application remained highly reliable (0.8%). The comprehensive error analysis revealed that 51.9% of failures occurred across multiple criteria simultaneously, highlighting the interconnected nature of legal reasoning components. This finding reveals significant hallucination risks, particularly in precedent citation where the system frequently generated non-existent case references despite correct legislative applications. Future work includes constructing more comprehensive evaluation datasets, addressing temporal dynamics in legal interpretation, and expanding the precedent database to include lower court rulings. The system demonstrates potential for supporting legal education and preliminary case analysis, though deployment in critical legal decision-making contexts requires additional safeguards and human oversight. The 61.7% accuracy rate, while promising, underscores the need for continued research before practical implementation.

7 Limitations

Our study has several limitations that should be acknowledged:

- **Limited Evaluation Dataset**: Our evaluation relied solely on question-answer pairs from legal manuals, lacking the diversity of real-world legal scenarios. A more comprehensive evaluation dataset encompassing varied legal contexts and query types would enable more thorough performance assessment.

- **Temporal Dynamics**: Legal interpretations change over time, affecting which laws and precedents are optimal for a given query. Our current framework lacks version control functionality to account for these temporal changes in legal content, potentially leading to outdated or superseded legal guidance.

- **Dataset Scope**: Our study faces significant limitations in both dataset scope and precedent coverage. The dataset scope is constrained to Korean criminal law cases derived from legal manuals, which may not represent the full spectrum of legal complexity encountered in practice. Additionally, our precedent database contains only Supreme Court cases, omitting lower court rulings that often provide relevant guidance for practical legal questions. This limited precedent coverage excludes district court decisions, appellate court rulings, and specialized court judgments that frequently address nuanced legal issues not covered by Supreme Court precedents. Future work should incorporate comprehensive multi-level court decisions to provide more complete legal coverage and expand dataset scope to include diverse legal domains.

- **Single Evaluator Bias**: The manual evaluation was conducted by a single author, potentially introducing subjective bias in accuracy assessments. Multi-evaluator scoring with inter-annotator agreement measures would strengthen the evaluation's reliability.

References

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Prof. 21st SIGIR*, page 335–336.

Jonathan H. Choi, Amy Monahan, and Daniel B. Schwarcz. 2023. Lawyering in the age of artificial intelligence. *109 Minnesota Law Review (Forthcoming 2024), Minnesota Legal Studies Research Paper*, pages 1–65.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: an expert-annotated NLP dataset for legal contract review. In *Prof. 35th NeurIPS*.

Jenny Hong, Derek Chong, and Christopher Manning. 2021. Learning from limited labels for long legal dialogue. In *Proc. Natural Legal Language Processing Workshop 2021*, pages 190–204.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Prof. 34th NIPS*, pages 9459 – 9474.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proc. Natural Legal Language Processing Workshop 2021*, page 19–35.

Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina-Athanasia Pantazi, and Manolis Koubarakis. 2021. Multi-granular legal topic classification on Greek legislation. In *Proc. Natural Legal Language Processing Workshop 2021*, page 63–75.

Bhawna Singh. 2024. Introduction to large language models. In *Building Applications with Large Language Models: Techniques, Implementation, and Applications*, pages 1–25. Apress.

Arianna Trozze, Toby Davies, and Bennett Kleinberg. 2024. Large language models in cryptocurrency securities cases: can a GPT model meaningfully assist lawyers? *Artifcial Intelligence and Law*.