KoBLEX: Open Legal Question Answering with Multi-hop Reasoning

Jihyung Lee [*1], **Daehui Kim** [*1,3], **Seonjeong Hwang**[1]
Hyounghun Kim[1,2], **Gary Geunbae Lee**[1,2]
[1]Graduate School of Artificial Intelligence, POSTECH, Republic of Korea
[2]Department of Computer Science and Engineering, POSTECH, Republic of Korea
[3]AI Future Lab, KT, Republic of Korea
{jihyung.lee, andrea0119, seonjeongh, h.kim, gblee}@postech.ac.kr

Abstract

Large Language Models (LLM) have achieved remarkable performances in general domains and are now extending into the expert domain of law. Several benchmarks have been proposed to evaluate LLMs' legal capabilities. However, these benchmarks fail to evaluate open-ended and provision-grounded Question Answering (QA). To address this, we introduce a **Ko**rean **B**enchmark for **L**egal **EX**plainable QA (**KoBLEX**), designed to evaluate provision-grounded, multi-hop legal reasoning. KoBLEX includes 226 scenario-based QA instances and their supporting provisions, created using a hybrid LLM–human expert pipeline. We also propose a method called **Par**ametric provision-guided **Se**lection **R**etrieval (**ParSeR**), which uses LLM-generated parametric provisions to guide legally grounded and reliable answers. ParSeR facilitates multi-hop reasoning on complex legal questions by generating parametric provisions and employing a three-stage sequential retrieval process. Furthermore, to better evaluate the legal fidelity of the generated answers, we propose **L**egal **F**idelity **Eval**uation (**LF-Eval**). LF-Eval is an automatic metric that jointly considers the question, answer, and supporting provisions and shows a high correlation with human judgments. Experimental results show that ParSeR consistently outperforms strong baselines, achieving the best results across multiple LLMs. Notably, compared to standard retrieval with GPT-4o, ParSeR achieves 37.91 higher F-1 and 30.81 higher LF-Eval. Further analyses reveal that ParSeR efficiently delivers consistent performance across reasoning depths, with ablations confirming the effectiveness of ParSeR. [1]

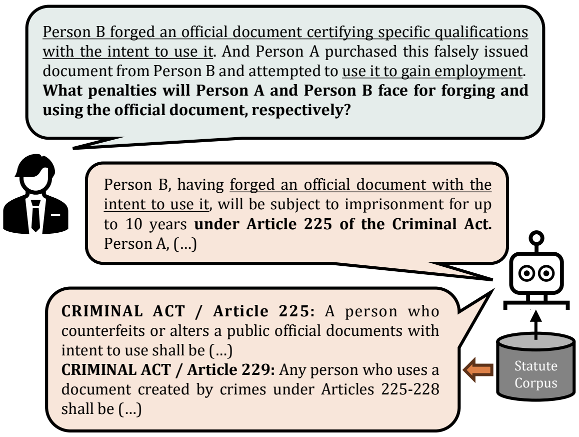
Person B forged an official document certifying specific qualifications with the intent to use it. And Person A purchased this falsely issued document from Person B and attempted to use it to gain employment. What penalties will Person A and Person B face for forging and using the official document, respectively?
Person B, having forged an official document with the intent to use it, will be subject to imprisonment for up to 10 years under Article 225 of the Criminal Act. Person A, (...)
CRIMINAL ACT / Article 225: A person who counterfeits or alters a public official documents with intent to use shall be (...) CRIMINAL ACT / Article 229: Any person who uses a document created by crimes under Articles 225-228 shall be (...)
Statute Corpus

Figure 1: **Overview of KoBLEX structure and task design.** Given a complex legal question, the system is required to reason over multiple statutory provisions.

1 Introduction

Large Language Models (LLM) have demonstrated strong performance across a wide range of tasks (Zhao et al., 2025), leading to the development of diverse benchmarks across general domains (Hendrycks et al., 2021; Cobbe et al., 2021; Zhou et al., 2023; Zheng et al., 2023; Rein et al., 2024). As LLMs increasingly demonstrate expert-level capabilities, interest in their application to the legal domain has grown, leading to the development of several legal benchmarks (Peng et al., 2023; Sun, 2023; Fei et al., 2024; Guha et al., 2023; Li et al., 2024b). Although recent legal benchmarks offer a diverse set of tasks to assess LLMs' capabilities in the legal domain, they are not well-suited for evaluating open-ended and provision-grounded legal question answering (QA) (Son et al., 2024; Kim et al., 2024). In practice, users often pose complex legal questions and expect answers grounded in legal provisions, as illustrated in Figure 1. Such grounding is critically essential in the legal domain, where hallucinated or inaccurate information can easily lead to serious situation (Engstrom and Gelbach, 2021; Romoser, 2023; Dahl et al., 2024). However, generating responses grounded in legal provisions is challenging because it requires not only identifying relevant provisions but also inter-

[*]Equal Contribution.

[1]The code and dataset are available at https://github.com/daehuikim/KoBLEX

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 4019–4053
November 4-9, 2025 ©2025 Association for Computational Linguistics

preting them with sufficient expert knowledge.

In light of these limitations, there is a clear need for a comprehensive evaluation of open-ended, provision-grounded legal QA. To this end, we present a **Ko**rean **B**enchmark for **L**egal **EX**plainable open-ended QA (KoBLEX), designed to evaluate multi-hop legal reasoning capabilities. KoBLEX comprises 226 multi-hop questions, answers, and their supporting statutory provisions, curated through a hybrid pipeline that combines LLM-based generation with expert revision and evaluation. While strict filtering pipeline leads to a limited sample size, experimental results show that KoBLEX serves as an effective benchmark for distinguishing reasoning capabilities across diverse methods. Moreover, unlike traditional legal benchmarks that rely on simple matching tasks or multiple-choice questions, KoBLEX evaluates methods' ability to generate free-form answers grounded in legal provisions. To promote accessibility and multilingual research, all instances are provided in both Korean and English. As illustrated in Figure 1, KoBLEX facilitates the evaluation of provision retrieval accuracy and multi-hop legal reasoning based on generated free-form answers.

Given the knowledge-intensive nature of provision-grounded legal QA, accurate retrieval is critical for generating factual answers. In light of this, we introduce **Par**ametric provision-guided **Se**lection **R**etrieval (ParSeR). ParSeR first generates *parametric provisions*, provisions constructed using the LLM's parametric knowledge to emulate the structure and language of real statutes. These serve as query scaffolds to improve the retrieval of relevant legal provisions. ParSeR then identifies the most relevant provision through a three-stage Retrieve–Rerank–Selection retrieval to answer the open-ended complex legal question.

To reliably assess whether the generated answers are faithful to the question, we also propose **L**egal **F**idelity **Eval**uation (LF-Eval). LF-Eval is built on the G-Eval (Liu et al., 2023), using instances from KoBLEX to assess legal fidelity. LF-Eval shows robust performance, achieving a Pearson correlation of 84.90 with human judgments.

Experimental results demonstrate that ParSeR outperforms strong retrieval-augmented reasoning baselines by consistently delivering performance gains across multiple LLMs and diverse evaluation metrics, including LF-Eval. Notably, with GPT-4o, ParSeR improves provision retrieval accuracy over one-time retrieval by +37.91 F-1 and +19.91 EM and enhances answer quality by +19.39 token-level F-1 and +30.81 LF-Eval, demonstrating its effectiveness. ParSeR consistently outperforms baselines across different reasoning depths. Further ablation analysis reveals that each component of ParSeR contributes to performance gains. In addition, ParSeR demonstrates greater efficiency compared to other baselines, achieving consistently strong results while generating much fewer tokens. Our contributions are summarized as follows:

- We introduce **KoBLEX**, a bilingual Korean-English benchmark of 226 provision-grounded, multi-hop legal QA instances curated via LLM and human validation pipeline.
- We introduce **ParSeR**, which combines LLM-generated parametric provisions with a three-stage retrieval pipeline, significantly outperforming existing retrieval-augmented reasoning baselines across multiple LLMs.
- We propose **LF-Eval**, a legal fidelity evaluation metric that excels at assessing the legal accuracy and provision alignment of generated responses.
- Experiments and analyses demonstrate the effectiveness and efficiency of ParSeR across models, metrics, and reasoning depths.

2 Related Works

Legal Benchmarks. A growing body of work has developed benchmarks for evaluating LLMs in legal contexts, spanning a diverse range of jurisdictions, languages, and task formats (Kim et al., 2016; Chalkidis et al., 2019; Zhong et al., 2020; Tuggener et al., 2020; Chalkidis et al., 2022; Kapoor et al., 2022; Chalkidis et al., 2023; Zhang et al., 2023; Li et al., 2024a). LegalBench (Guha et al., 2023) proposes 162 few-shot tasks in English across six categories of legal reasoning, offering one of the most comprehensive benchmark suites to date. LawBench (Fei et al., 2024) adapts this paradigm to Chinese law, comprising 20 tasks that assess capabilities in judgment prediction, statutory interpretation, and legal knowledge retrieval.

In the Korean legal domain, several benchmarks have been introduced to evaluate domain-specific language understanding. LBOX OPEN (Hwang et al., 2022) presents a large-scale multi-task benchmark constructed from Korean court decisions, encompassing tasks such as classification and judgment prediction. KMMLU (Son et al., 2024) includes Korean legal QA as a multiple-choice cate-

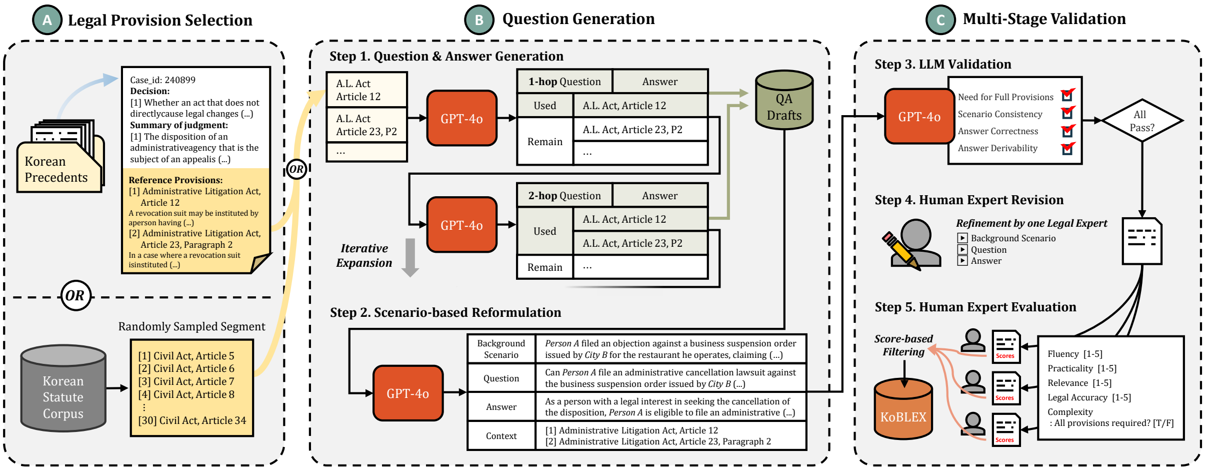
A
Legal Provision Selection
Korean Precedents
Case_id: 240899
Decision:
[1] Whether an act that does not directlycause legal changes (...)
Summary of judgment:
[1] The disposition of an administrativeagency that is the subject of an appealis (...)
Reference Provisions:
[1] Administrative Litigation Act, Article 12
A revocation suit may be instituted by aperson having (...)
[2] Administrative Litigation Act, Article 23, Paragraph 2
In a case where a revocation suit isinstituted (...)
OR
OR
Randomly Sampled Segment
Korean Statute Corpus
[1] Civil Act, Article 5
[2] Civil Act, Article 6
[3] Civil Act, Article 7
[4] Civil Act, Article 8
⋮
[30] Civil Act, Article 34
B
Question Generation
Step 1. Question & Answer Generation
A.L. Act Article 12
A.L. Act Article 23, P2
...
GPT-4o
1-hop Question
Answer
Used
A.L. Act, Article 12
Remain
A.L. Act, Article 23, P2
...
QA Drafts
GPT-4o
2-hop Question
Answer
Used
A.L. Act, Article 12
A.L. Act, Article 23, P2
Remain
...
Iterative Expansion
Step 2. Scenario-based Reformulation
GPT-4o
Background Scenario
Person A filed an objection against a business suspension order issued by City B for the restaurant he operates, claiming (...)
Question
Can Person A file an administrative cancellation lawsuit against the business suspension order issued by City B (...)
Answer
As a person with a legal interest in seeking the cancellation of the disposition, Person A is eligible to file an administrative (...)
Context
[1] Administrative Litigation Act, Article 12
[2] Administrative Litigation Act, Article 23, Paragraph 2
C
Multi-Stage Validation
Step 3. LLM Validation
GPT-4o
Need for Full Provisions
Scenario Consistency
Answer Correctness
Answer Derivability
All Pass?
Step 4. Human Expert Revision
Refinement by one Legal Expert
Background Scenario
Question
Answer
Step 5. Human Expert Evaluation
Score-based Filtering
KoBLEX
Scores
Scores
Scores
Fluency [1-5]
Practicality [1-5]
Relevance [1-5]
Legal Accuracy [1-5]
Complexity
: All provisions required? [T/F]

Figure 2: **Generation and validation pipeline for KoBLEX.** The pipeline consists of three stages: (A) context construction from either randomly sampled segment of statute corpus or reference provisions in precedents, (B) question–answer generation using GPT-4o based on the selected legal context, and (C) multi-stage validation. An initial LLM-based evaluation filters out incomplete or unsupported pairs based on predefined criteria. Subsequently, one legal expert manually revised each validated pair and three legal experts rated them along five dimensions: fluency, practicality, relevance, legal accuracy, and complexity.

gory within a broader zero-shot evaluation suite. KBL (Kim et al., 2024) offers a diverse set of multiple-choice legal tasks, including bar exam and scenario-based questions.

However, all of these benchmarks share key limitations: they rely on multiple-choice or binary formats, lack explicit links to relevant statutory provisions, and do not support open-ended QA. As a result, they are not suitable for evaluating models' ability to produce explainable and factually grounded answers in complex legal settings. These limitations motivate the development of KoBLEX, a Korean benchmark for open-ended legal QA that requires provision-grounded, open-ended and multi-hop reasoning over statutory provisions.

Retrieval Augmented Reasoning. Standard Prompting (SP) was proposed to show that LLMs can leverage their parametric knowledge through in-context learning (Brown et al., 2020). Building on this, Chain of Thought (CoT) enables step-by-step reasoning by explicitly activating the model's parametric knowledge (Wei et al., 2022). However, LLMs often fail to align with knowledge-intensive factual information, leading to hallucination (Huang et al., 2025). To address this limitation, researchers have increasingly explored Retrieval-Augmented Generation (RAG), which enhances response accuracy by incorporating relevant external knowledge (Lewis et al., 2020; Gao et al., 2024; Cho and Lee, 2025). While RAG leverages external knowledge, it often fails to integrate external knowledge with the model's parametric knowledge, limiting its performance on complex reasoning tasks. To overcome this, recent work has introduced Retrieval-Augmented Reasoning (RARE), which aims to combine retrieval with multi-step reasoning capabilities. Self-Ask (Press et al., 2023) introduces iterative RARE by generating intermediate questions and querying external knowledge to derive the final answer. IRCoT (Trivedi et al., 2023) interleaves CoT traces with retrieval, using the interleaved generations to incorporate external knowledge into the reasoning process. FLARE (Jiang et al., 2023) improves upon IRCoT by using the model's token-level confidence to adaptively retrieve contexts. ProbTree (Cao et al., 2023) decomposes complex questions into a tree of sub-queries, solves each node using diverse strategies, and aggregates the results based on token-level confidence to produce the final answer. BeamAggr (Chu et al., 2024) enhances ProbTree by performing multi-source reasoning to generate answer candidates at leaf nodes, followed by beam combination and probabilistic answer aggregation. Despite advances in complex multi-hop reasoning, these approaches remain underexplored in knowledge-intensive domains like law. This underscores the need for methods that effectively handle complex multi-hop reasoning in knowledge-intensive domains. To bridge this gap, our research propose ParSeR, a novel framework that integrates the parametric knowledge of LLMs with a 3-stage retrieval pipeline to support effective retrieval-augmented reasoning.

Background Scenario	Person B forged an official document certifying specific qualifications with the intent to use it. Person A purchased this falsely issued document from Person B and attempted to use it to gain employment.
Question	What penalties will Person A and Person B face for forging and using the official document, respectively?
Answer	Person B, having forged an official document with the intent to use it, will be subject to imprisonment for up to 10 years under Article 225 of the Criminal Act. Person A, having used a document created in violation of Article 225, will be subject to imprisonment for up to 10 years under Article 229 of the Criminal Act.
Reference Provision [1]	**CRIMINAL ACT / Article 225**: A person who counterfeits or alters a public official document with intent to use shall be punished by imprisonment with labor for not more than ten years.
Reference Provision [2]	**CRIMINAL ACT / Article 229**: Any person who uses a document created by crimes under Articles 225–228 shall be punished by the penalty prescribed for such crimes.

Figure 3: **Example QA instance from the KoBLEX (translated from Korean).** This multi-hop question requires interpreting multiple statutes. Yellow texts highlight key legal information essential for deriving the correct answer.

3 KoBLEX

In this section, we describe the construction process of KoBLEX. We first generate initial drafts using an LLM, then filter and revise them through a multi-stage validation process involving both LLM-based and human expert review (Figure 2).

3.1 Legal Provision Selection

Korean statutes are systematically structured at the article and paragraph level, and adjacent provisions typically exhibit strong topic continuity by addressing the similar legal concept or regulatory subject (Ministry of Government Legislation, 2024). Accordingly, sampling a continuous segment is a reasonable heuristic for selecting relevant legal content. However, this approach may limit the diversity of questions spanning multiple legal sources.

To complement this approach, we also extract statutes cited in real-world court decisions. Specifically, we utilize the *reference provisions* field in Korean precedents to identify statutory clauses that are actually invoked in judicial reasoning. Data sources are described in Appendix C.

3.2 Question Generation

Question & Answer Generation. Based on the statutory provisions selected in the previous step, we use GPT-4o (OpenAI et al., 2024) to generate initial drafts of question-answer pairs.

As illustrated in Step 1 of Figure 2, we adapt an incremental generation strategy. We first prompt the model to generate single-hop questions, each of which can be answered using only one provision. Then, based on these single-hop questions, we incrementally expand them into two-hop and three-hop versions by introducing additional provisions and guiding the model to integrate them logically into the reasoning process.

Scenario-based Reformulation. To better reflect realistic legal situations, we reformulate each question–answer pair into a fact-based legal scenario using GPT-4o. This transformation encourages more natural multi-hop reasoning and enhances the dataset's alignment with practical legal contexts. Accordingly, each QA draft is rewritten as a scenario-based item structured as realistic legal facts and anonymized parties (e.g., Person A).

After this process, each QA instance is structured into four components: Background Scenario B, Question Q, Answer A, Context C, where $C = \{p_1, p_2, \ldots, p_n\}$ denotes the set of reference statutory provisions used to support the reasoning. A representative QA instance is shown in Figure 3.

3.3 Multi-Stage Validation

LLM Validation. Inspired by prior work that uses LLMs as evaluators for quality control (Bedi et al., 2024), we employ GPT-4o to filter incorrect QA samples before the human validation step. In this step, as illustrated in Step 3 of Figure 2, the LLM automatically evaluates whether each question Q requires all C, and the consistency of B, Q, and A. We term each step as *Partial Check* and *Full Check*. *Partial Check* is designed to ensure that Q genuinely requires the full set of C for resolution. Given $C = \{p_1, p_2, \ldots, p_n\}$, we evaluate whether any non-empty subset of the powerset of C is sufficient to answer Q. If any such subset yields a correct answer without referencing the remaining provisions, we consider the instance to lack true multi-hop characteristics and exclude it. *Full Check* involves a comprehensive validation of the triplet (B, Q, A) through an inclusive single evaluation prompt, covering the following three aspects: (1) *Scenario consistency*, which assesses whether

		# of words		
	# Examples	**Background scenario**	**Question**	**Answer**
1-hop	55	33.96	13.35	11.67
2-hop	125	37.45	18.78	29.56
3-hop	46	46.2	27.13	53.11
Total	226	38.38	19.16	30.0

Table 1: **Statistics on KoBLEX.** The average numbers of words on each QA component are categorized by reasoning depth (1-hop, 2-hop, and 3-hop).

(B, Q) is logically and legally coherent with C; (2) *Answer correctness*, which determines whether A aligns with the statutory interpretation of C; and (3) *Answer derivability*, which checks whether A can be fully inferred from C without requiring any unstated assumptions. The prompt templates used for both checks are provided in Appendix F.

Human Expert Revision. To ensure legal correctness, clarity, and linguistic fluency, QA instances are further revised and verified by Korean law school graduates and students. Each QA instance is classified as either *Pass*, *Revise*, or *Hold*. Instances marked as *Revise*—such as ambiguous legal actors in the scenario or legally incorrect answers—are corrected accordingly. In the case of *Hold*, most instances are excluded from the final dataset unless the issue can be resolved by appending additional statutory provisions. Details about human expert revision are in Appendix E.

Human Expert Evaluation. Following the revision, each QA instance is evaluated by three different legal experts using five criteria: *Fluency*, *Practicality*, *Relevance*, *Legal Accuracy*, and *Complexity*. Fluency, practicality, relevance, and legal accuracy are rated on a 5-point Likert scale, while complexity is a binary label (True/False), with True indicating that the question requires utilizing the full set of C. Descriptions of each criterion and the evaluation guidelines are provided in Appendix E.

Instances that show no critical issues during revision and received sufficiently high evaluation scores across these criteria are curated for inclusion in the final dataset. Details of the filtering process are provided in Appendix A.1.

3.4 Data statistics

KoBLEX comprises 226 high-quality QA instances. Table 1 summarizes the number of examples and the average word count for each QA component grouped by reasoning depth. As reasoning depth increases, the average word count consistently rises across all components, clearly reflecting the greater linguistic and logical complexity involved in multi-hop legal reasoning.

The final benchmark spans 83 distinct Korean statutes, including major codes such as the *Civil Act*, *Criminal Act*, and *Criminal Procedure Act*. A complete list of statutes and their distribution is provided in Appendix B. Among the 226 QA instances, 153 are generated using reference provisions extracted from Korean court decisions, while the remaining 73 are constructed from randomly sampled segments of the Korean statute corpus.

The overall average scores across the five evaluation dimensions are: *Fluency* 4.385, *Practicality* 4.435, *Relevance* 4.540, *Legal Accuracy* 4.515, and *Complexity* 0.915. The agreement between the evaluators also reaches 96%, with at least two of the three experts assigning consistent labels. Details about inter-annotator agreement are in Appendix D.

While fluency and practicality may involve some degree of subjectivity and are less directly tied to legal reasoning, the consistently high scores in relevance, legal accuracy, and complexity suggest that the dataset effectively captures the demands of factually grounded, multi-hop legal reasoning. Appendix A.2 summarizes the data filtered at each LLM and human validation stage.

3.5 English Version of KoBLEX

Unlike previous Korean legal benchmarks (Hwang et al., 2022; Kim et al., 2024), which only provide Koean data, we release an English-translated version of KoBLEX to promote broader accessibility and facilitate multilingual research on legal question answering. Specifically, B, Q, and A are translated using GPT-4o, while the statutory provisions C are primarily based on high-quality official translations provided by the Korea Legislation Research Institute. [2] A small subset of provisions in C that are not covered by the official source are translated via GPT-4o and marked with the tag *%MACHINE_TRANSLATED%*.

4 ParSeR

We introduce Parametric provision guided Selection Retrieval (ParSeR), a method designed to retrieve supporting legal provisions for complex legal questions effectively.

[2] https://elaw.klri.re.kr/eng_service/main.do

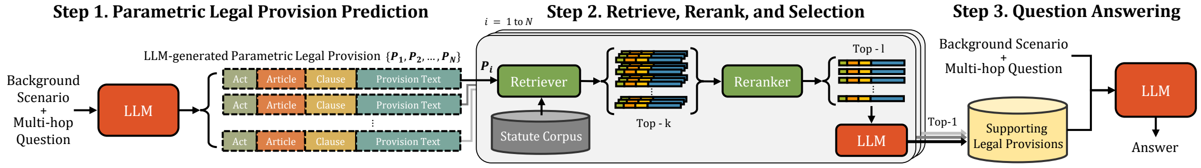
Step 1. Parametric Legal Provision Prediction
Step 2. Retrieve, Rerank, and Selection
Step 3. Question Answering
i = 1 to N
LLM-generated Parametric Legal Provision $\{P_1, P_2, \ldots, P_N\}$
Background Scenario + Multi-hop Question
LLM
Act
Article
Clause
Provision Text
P_i
Retriever
Statute Corpus
Top - k
Reranker
Top - l
LLM
Top-1
Background Scenario + Multi-hop Question
Supporting Legal Provisions
LLM
Answer

Figure 4: **Illustration of PARSER.** (Step 1) The LLM initially generates parametric provisions. (Step 2) Parametric provisions are then used as queries for *Retrieve, Rerank, and Selection* retrieval. (Step 3) Finally, collected supporting legal provisions are used to multi-hop reasoning for generating responses.

4.1 Parametric Provision Generation

Complex legal questions (Q) often require reasoning over multiple statutory provisions. However, relying solely on the parametric knowledge of LLMs can be unreliable, as they are prone to hallucinations and may fail to recall the exact legal texts (Dahl et al., 2024). To address this, we incorporate retrieval over an actual statute corpus to ground the model's responses in authoritative legal text. As illustrated in Step 1 of Figure 4, we instruct the LLM to generate a set of potentially relevant provisions ($\{p_n{}_{n=1}^{N}\}$) from multi-hop legal question (Q), where N is the number of generated parametric provisions. Each p_n reflects a distinct statutory component that may support reasoning over Q. Since $\{p_n{}_{n=1}^{N}\}$ are generated solely based on the LLM's parametric knowledge, we refer to them as parametric provisions and use them only as intermediate queries. This allows us to leverage retrieval based on queries that resemble the target question.

4.2 Retrieve, Rerank and Selection

To improve retrieval accuracy, we propose a novel three-stage retrieval approach incorporating Bi-encoder (Huang et al., 2013) retrieval, Cross-encoder (Reimers and Gurevych, 2019) reranking, and selection via an LLM. As illustrated in Step 2 of Figure 4, each parametric provision p_n is used to retrieve the top k most relevant statutory provisions ($Top - k$) from the corpus via a Bi-encoder retriever based on cosine similarity. The $Top - k$ provisions retrieved for each p_n are reranked using a Cross-encoder reranker to better capture fine-grained relevance. After reranking, we select $Top - l$, $(l < k)$ provisions from $Top - k$ provisions. Then, we instruct an LLM to select the most relevant one among the $Top - l$ for each p_n. This process enables the collection of reliable supporting legal provisions by leveraging generated parametric provisions. Finally, the collected supporting legal provisions are fed into the LLM to solve the complex multi-hop legal question through provision-grounded reasoning. The parametric provisions generated in the initial stage facilitate multi-hop reasoning by enabling PARSER to search for each piece of supporting provision.

Legal Fidelity Evaluation Prompt

<Task Description>
You will be given a complex legal question along with the relevant legal provisions that can be used to resolve it.

<Evaluation Criteria>
Evaluate the legal accuracy of the response on a scale 1 $\sim$ 10.

<Evaluation Steps>
1. Check whether the prediction properly answers the question.
2. Check whether the prediction contradicts or omits any legal provisions in the context.
3. Heavily penalize when the legal conclusion differs in detail from the expected output.
4. Heavily penalize if the prediction contradicts or omits any specific elements from the context.
5. Heavily penalize responses that include statements like 'The given context does not include the answer, but generally'.

<Query>
Question: {question} ; Context: {context}
Expected output: {answer} ; Prediction: {prediction}

Figure 5: **Prompt of Legal Fidelity Evaluation (LF-EVAL).** {placeholder} indicates a slot to be filled with the corresponding value for evaluation.

5 LF-EVAL

KOBLEX is a benchmark designed to evaluate the legal fidelity of responses generated by LLMs. However, existing evaluation metrics often fail to assess the legal correctness of these responses (Trautmann et al., 2024). To address this, we introduce Legal Fidelity Evaluation (LF-EVAL), an evaluation framework for measuring legal fidelity.

LF-EVAL builds upon G-Eval (Liu et al., 2023), a representative LLM-as-a-Judge evaluation approach. It evaluates the legal fidelity of a response by comparing it against the reference legal provisions and the expected answer. Figure 5 illustrates the prompt used in LF-EVAL. Under the task description of answering legal questions, LF-EVAL employs a robust LLM judge to assign a score on a

	F-1			EM			Token F-1			LF-EVAL		
	Qwen	EXAONE	GPT-4o	Qwen	EXAONE	GPT-4o	Qwen	EXAONE	GPT-4o	Qwen	EXAONE	GPT-4o
SP$^{\diamondsuit}$ (Brown et al., 2020)	-	-	-	-	-	-	30.52	23.04	36.20	40.74	49.46	55.00
CoT$^{\diamondsuit}$ (Wei et al., 2022)	-	-	-	-	-	-	26.37	23.71	32.34	39.41	37.42	52.75
SP (Brown et al., 2020) + OR$^{\heartsuit}$	21.50	21.50	21.50	7.08	7.08	7.08	32.18	20.58	26.75	45.36	43.98	36.45
CoT (Wei et al., 2022) + OR$^{\heartsuit}$	21.50	21.50	21.50	7.08	7.08	7.08	28.39	28.82	30.68	42.10	47.26	46.42
Self-Ask$^{\spadesuit}$ (Press et al., 2023)	9.29	8.55	8.55	2.65	1.33	1.77	16.59	14.72	7.82	34.14	37.77	22.44
IRCoT$^{\spadesuit}$ (Trivedi et al., 2023)	20.42	15.89	23.91	4.42	1.77	4.42	31.62	26.31	31.39	46.78	48.20	46.68
FLARE$^{\spadesuit}$ (Jiang et al., 2023)	40.64	25.23	31.75	3.98	14.16	4.42	29.54	21.31	34.37	53.76	34.66	50.55
ProbTree$^{\clubsuit}$ (Cao et al., 2023)	15.84	11.61	17.32	2.65	2.21	3.98	28.38	24.67	33.91	43.77	46.74	52.62
BeamAggr$^{\clubsuit}$ (Chu et al., 2024)	14.05	10.83	16.89	2.65	0.44	3.54	16.02	10.83	22.83	32.31	31.46	41.59
PARSER$^{\clubsuit}$ (Ours)	**46.24**	**48.73**	**59.41**	**17.70**	**17.70**	**26.99**	**40.65**	**31.09**	**46.14**	**56.00**	**57.58**	**67.26**

Table 2: **Experimental results of various baseline methods on KOBLEX.** Columns shaded in blue measure retrieval accuracy, and columns shaded in yellow measure generation accuracy. The best results are highlighted in bold. We utilize Qwen3-32B (Team, 2025), EXAONE-3.5-32B (Research, 2024), and GPT-4o (OpenAI et al., 2024). (◇: No-retrieval, ♡: One-time retrieval, ♠: Iterative retrieval, ♣: Sub-query retrieval).

1–10 scale. The evaluation follows five clearly defined steps, each corresponding to specific criteria: *Answer Relevance*, *Legal Consistency*, *Conclusion Accuracy*, *Context Fidelity*, and *Avoid Generic Responses*. Finally, LF-EVAL produces both a scalar score and detailed justifications aligned with each of the five evaluation steps. These suggest that LF-EVAL not only provides a reliable score, but also offers interpretable explanations.

To assess LF-EVAL's reliability, we conduct a human evaluation study on generated responses from KOBLEX, using two independent annotator groups. The results show that LF-EVAL achieves a Pearson correlation of 84.90 with human judgments, outperforming existing evaluation metrics. Examples and details of the human evaluation study are provided in the Appendix G.

6 Experiments

We describe our experimental setup, including models, evaluation metrics, retrieval setting, and baselines. Additional details are in Appendix H.

6.1 Models

We employ three different LLMs: Qwen3 (Team, 2025), EXAONE-3.5 (Research, 2024), and GPT-4o (OpenAI et al., 2024). These models are selected for their strong Korean language understanding capabilities. We use BM-25 (Robertson and Zaragoza, 2009) as the retriever and BGE (Chen et al., 2024), finetuned on the Korean dataset, as the reranker.

6.2 Metrics

Since KOBLEX includes gold-supporting provisions and answers, we employ retrieval and generation metrics for comprehensive evaluation. For retrieval performance, we use Exact Match (EM) and F-1, computed by comparing the retrieved provisions against the gold-supporting provisions. For generation performance, we report Token F-1, a standard metric for evaluating free-form QA, and LF-EVAL for assessing the legal fidelity.

6.3 Statute Corpus

We construct a paragraph-level statute corpus to provide fine-grained legal information. We include every active statute that has been cited in Korean court decisions between 1998 and 2024[3]. The final corpus comprises 608 unique statutes, totaling about 233,544 paragraph-level provisions. We use this statute corpus as a retrieval pool to obtain actual provisions for our experiments.

6.4 Baselines

To evaluate the effectiveness of PARSER, we compare it against a diverse set of multi-hop reasoning baselines. Since many baselines target general-domain tasks, official implementations are incompatible for KOBLEX. Therefore, we manually reimplemented them for fair comparison. We perform retrieval over our statute corpus to ensure consistency for methods requiring retrieval.

We include the following methods as baselines: Standard Prompting (SP) (Brown et al., 2020), Chain of Thought (CoT) (Wei et al., 2022), Self-Ask (Press et al., 2023), IRCoT (Trivedi et al., 2023), FLARE (Jiang et al., 2023), ProbTree (Cao et al., 2023), and BeamAggr (Chu et al., 2024). We also include a simple One-Time Retrieval baseline, which gives oracle access to the same number of gold reference provisions as the required number of reasoning hops. Implementation details about baseline methods are provided in Appendix H.3

[3] We obtain documents via the official API (`https://open.law.go.kr/`) on March 17, 2025.

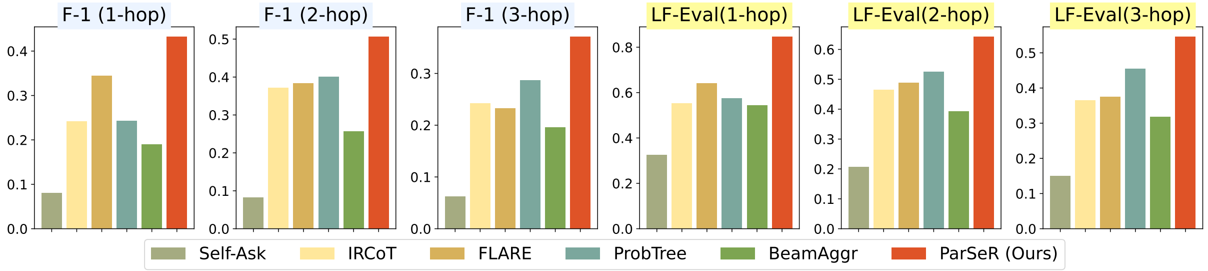

Figure 6: Retrieval performance (F-1) and generation performance (LF-EVAL) of each method by reasoning depth.

	F-1	EM	Token F-1	LF-EVAL
PARSER	**48.74**	**17.70**	**31.09**	**57.58**
w/o Selection	40.61	13.72	27.11	50.18
w/o Reranking	40.64	14.16	29.54	54.02
w/o Reranking, Selection	27.56	6.64	25.90	45.97
w/o Reranking, Selection, Provision	21.41	3.98	21.20	45.52

Table 3: **Ablation results on PARSER on EXAONE.** "w/o Provision" replaces parametric provision generation with top-k retrieval based on the original question, where k matches the number of generated provisions.

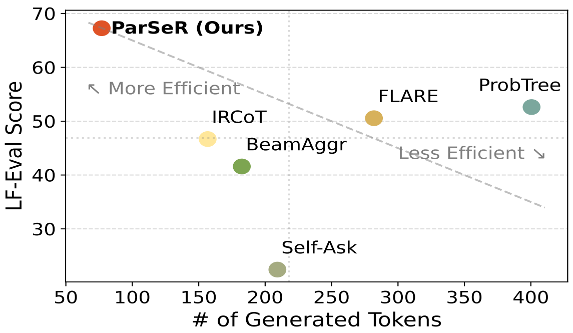

Figure 7: Average PARSER scores with respect to average number of generated tokens by GPT-4o.

7 Results

Table 2 shows experimental results on KOBLEX, comparing various baseline methods. We observe that baselines with low F-1 and EM scores tend to show lower Token F-1 and LF-EVAL scores than the No-retrieval baseline.[4] This suggests that retrieving irrelevant provisions negatively impacts multi-hop legal reasoning.

On the other hand, PARSER consistently outperforms all baseline methods with all LLMs across all retrieval and generation metrics. Notably, on GPT-4o, PARSER significantly surpasses the strongest baseline, ProbTree, with a +12.23 improvement in Token F-1 and +14.64 in LF-EVAL. This indicates that PARSER can effectively perform reasoning over complex multi-hop legal questions. Additional results on smaller LLMs are in Appendix J.

8 Analyses

Reasoning Depth. Given that KOBLEX encompasses scenario-based QA tasks requiring varying levels of reasoning depth, we conduct a detailed analysis by evaluating performance across different hop levels. Figure 6 shows the performances with respect to different reasoning depths. PARSER consistently achieves the best performance across all hop levels. Excluding PARSER, confidence-based methods such as FLARE and ProbTree outperform other baselines. Specifically, FLARE ranks second at the 1-hop level, while ProbTree excels in deeper reasoning at the 2-hop and 3-hop levels. These results suggest that the performance of different existing methods can vary depending on the required reasoning depth. In contrast, PARSER demonstrates robust performance across questions with all levels of reasoning depth.

Ablation Study. To analyze the role of each module in driving the performance improvements of PARSER, we conduct an ablation study by isolating the impact of each component. Table 3 shows the ablation study of our method on EXAONE 3.5-32B (Research, 2024). Removing *Selection* causes a greater performance drop than removing *Reranking*, suggesting that leveraging the LLM's ability to select relevant provisions is more effective than cross-encoder. We observe a significant performance drop when both *Reranking* and *Selection* are removed. Furthermore, replacing *parametric provision* generation with simple top-k retrieval based on the original question, where k equals the number of generated provisions, leads to even worse performance. This indicates that parametric provision generation better supports multi-hop retrieval-augmented reasoning than simple retrieval.

[4] While most baselines generally improve with iterative reasoning, BeamAggr underperforms because its multi-source retrieval is incompatible with our setup, limiting its advantage.

	F-1	EM	Token F-1	LF-EVAL
PARSER (Sparse)	48.74	**17.70**	**31.09**	**57.58**
PARSER (Dense)	**50.43**	16.81	29.56	**57.58**
PARSER (Hybrid)	48.16	16.81	28.76	56.61

Table 4: **Performance of PARSER across different retriever types.** We compare sparse (BM25 (Robertson and Zaragoza, 2009)), dense (BGE-M3 (Chen et al., 2024)), and hybrid (BGE-M3 hybrid) retrievers using EXAONE 3.5-32B (Research, 2024).

Efficiency. While retrieval-augmented reasoning methods significantly improve performance on complex questions, they often require a large number of tokens for final answer prediction. To assess both effectiveness and efficiency, we analyze the performance of each method relative to the number of tokens consumed during answer prediction. Figure 7 presents LF-EVAL score against the average number of generated tokens. While methods with more generated tokens tend to show increased performance, the improvements are generally marginal relative to the computational overhead, suggesting potential inefficiency. However, PARSER demonstrates the highest performance while generating the fewest tokens, making it the most efficient approach. Unlike other baselines, PARSER leverages parametric provisions generated in the initial stage to facilitate targeted multi-hop evidence retrieval while focusing on improving retrieval accuracy through a 3-stage retrieval pipeline. This feature leads that PARSER is not only effective but also cost-efficient for multi-hop legal reasoning.

Effect of Retreiver Type. Since various types of retrieval tools are available, we investigate the impact of retriever type on the performance of PARSER. For the legal provision retrieval task, we consider BM25 (Robertson and Zaragoza, 2009) as a sparse retriever, BGE-M3 (Chen et al., 2024), known for its strong Korean embedding capabilities, as a dense retriever, and their combination as a hybrid retriever. Table 4 presents the experimental results of using different retriever types with EXAONE-3.5-32B. Overall, there is no significant performance difference across retriever types, although the sparse retriever (BM25) achieves the highest EM, token-level F1, and LF-EVAL score. Therefore, we adopt the sparse retriever (BM25) as the retrieval tool, considering its accessibility for reproducibility, efficient retrieval speed, and fairness in comparison. See Appendix H.3 for a detailed description of the retrieval configuration.

$Top-k$	F-1	EM	Token F-1	LF-EVAL
Top-50	47.98	16.81	30.24	58.93
Top-100	48.74	17.70	31.09	57.58
Top-200	49.29	17.70	31.20	**60.59**
Top-300	**50.26**	**18.58**	**31.97**	60.58

Table 5: **Analysis on k value on retrieval.** We vary the k value (50, 100, 200, 300) while keeping other modules of PARSER fixed, using EXAONE 3.5-32B.

$Top-l$	F-1	EM	Token F-1	LF-EVAL
Top-5	47.56	17.26	30.63	**59.21**
Top-10	48.74	17.70	**31.09**	57.58
Top-20	50.03	18.14	30.67	59.14
Top-30	**50.34**	**19.03**	30.43	58.49

Table 6: **Analysis on l value on reranking.** We vary the l value (5, 10, 20, 30) while keeping other modules of PARSER fixed, using EXAONE 3.5-32B.

Effect of k and l in PARSER. We investigate the impact of two key hyperparameters in the 3-stage retrieval pipeline: k (retrieval scope) and l (reranking scope). Table 5 presents the results of varying k while keeping other components of PARSER fixed. We observe a slight improvement in retrieval metrics as k increases, while the effect on generation metrics remains marginal. Table 6 shows the results of varying l under the same conditions. Unlike k, increasing l does not lead to consistent performance gains. While larger k and l values offer broader context and better recall, they also increase computational cost in reranking and risk exceeding context limits during selection. Therefore, we set $k = 100$ and $l = 10$ for PARSER in this paper.

9 Conclusion

In this work, we introduce **KOBLEX**, a benchmark designed to evaluate LLMs on provision-grounded, open-ended legal QA in Korean law. Furthermore, we present **PARSER**, which significantly outperforms existing retrieval-augmented reasoning methods in both retrieval accuracy and answer quality. To enable reliable assessment of legal fidelity in free-form responses, we propose **LF-EVAL**, an automatic evaluation aligned with human judgment.

Our experiments demonstrate that PARSER consistently outperforms all baselines across all metrics in a cost-efficient and reasoning-depth-agnostic manner. We hope KOBLEX, LF-EVAL, and PARSER serve as valuable resources for advancing research in legal NLP.

Limitations

Limited Scale and Expert Dependency. The current version of KoBLEX is relatively limited in size. While our automated pipeline enables the generation of initial QA drafts at scale, each instance still requires careful review and revision by legal experts to ensure legal correctness and contextual coherence. This expert validation step remains essential given the current limitations of LLMs in reliably handling nuanced legal interpretation without human oversight.

Civil Law Focus. KoBLEX, ParSeR, and LF-Eval are developed based on Korean statutory law, reflecting the characteristics of a civil law system where codified statutes serve as the primary source of legal authority. While this design enables rigorous evaluation of statute-grounded legal reasoning, it may limit the applicability to common law jurisdictions, such as those in the United States or the United Kingdom, where legal interpretation heavily relies on case law and judicial precedents.

Ethical Considerations

To ensure ethical data construction and usage, several precautions are taken during the development of KoBLEX.

Dataset Construction. First, all question, answer instances in the dataset are derived from fictional legal scenarios and do not contain any personally identifiable information. Every background scenario is composed using anonymized character names (e.g., Person A, Person B), and no real individuals, cases, or sensitive details are included.

Second, all legal experts who participated in the revision and evaluation process were informed in advance that the purpose of the task is to construct a dataset for evaluating LLM performance on legal reasoning (see Appendix E).

Finally, all legal documents used in constructing KoBLEX—both statutes and precedents—were collected from official government APIs and fall under Korea's public data policy, which permits their redistribution for research purposes. For the English versions of the statutes, we used publicly released translations by the Korea Legislation Research Institute (KLRI), which confirms redistribution is allowed for non-commercial research use. Any machine-translated content is labeled within the dataset to maintain transparency (Appendix C).

Intended Use. This work presents a benchmark and methodology for evaluating open-ended, provision-grounded legal question answering in Korean. It is designed solely for research purposes and is not intended for direct use in real-world legal decision-making or as a substitute for professional legal advice. The benchmark aims to support the development and evaluation of legal NLP systems.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2025-RS-2020-II201789) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation, Contribution Rate: 47.5%). This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (Project Name: Development of an AI-Based Korean Diagnostic System for Efficient Korean Speaking Learning by Foreigners, Project Number: RS-2025-02413038, Contribution Rate: 47.5%). This work was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH), Contribution Rate: 5%).

References

Suhana Bedi, Scott L Fleming, Chia-Chun Chiang, Keith Morse, Aswathi Kumar, Birju Patel, Jenelle A Jindal, Conor Davenport, Craig Yamaguchi, and Nigam H Shah. 2024. Quest-ai: A system for question generation, verification, and refinement using ai for usmle-style exams. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 54–69. World Scientific.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Shulin Cao, Jiajie Zhang, Jiaxin Shi, Xin Lv, Zijun Yao, Qi Tian, Lei Hou, and Juanzi Li. 2023. Probabilistic tree-of-thought reasoning for answering knowledge-intensive complex questions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12541–12560, Singapore. Association for Computational Linguistics.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Nicolas Garneau, Cătălina Goanță, Daniel Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jeonghun Cho and Gary Lee. 2025. K-COMP: Retrieval-augmented medical domain question answering with knowledge-injected compressor. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6878–6901, Albuquerque, New Mexico. Association for Computational Linguistics.

Zheng Chu, Jingchang Chen, Qianglong Chen, Haotian Wang, Kun Zhu, Xiyuan Du, Weijiang Yu, Ming Liu, and Bing Qin. 2024. BeamAggR: Beam aggregation reasoning over multi-source knowledge for multi-hop question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1229–1248, Bangkok, Thailand. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

David Freeman Engstrom and Jonah B Gelbach. 2021. Legal tech, civil procedure, and the future of adversarialism. *University of Pennsylvania Law Review*, pages 1001–1099.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Matthew Finlayson, John Hewitt, Alexander Koller, Swabha Swayamdipta, and Ashish Sabharwal. 2024. Closing the curious case of neural text degeneration. In *The Twelfth International Conference on Learning Representations*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.

Jeffrey Ip and Kritin Vongthongsri. 2025. deepeval.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Arnav Kapoor, Mudit Dhawan, Anmol Goel, Arjun T H, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, and Ashutosh Modi. 2022. HLDC: Hindi legal documents corpus. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3521–3536, Dublin, Ireland. Association for Computational Linguistics.

Mi-Young Kim, Randy Goebel, Yoshinobu Kano, and Ken Satoh. 2016. Coliee-2016: evaluation of the competition on legal information extraction and entailment. In *International Workshop on Juris-informatics (JURISIN 2016)*.

Yeeun Kim, Youngrok Choi, Eunkyung Choi, JinHwan Choi, Hai Jin Park, and Wonseok Hwang. 2024. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5573–5595, Miami, Florida, USA. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, et al. 2024a. Legalagentbench: Evaluating llm agents in legal domain. *arXiv preprint arXiv:2412.17259*.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024b. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *arXiv preprint arXiv:2409.20288*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. *Preprint*, arXiv:2407.03618.

Ministry of Government Legislation. 2024. Legislative drafting review guidelines. `https://www.moleg.go.kr/menu.es?mid=a10105030000`. Cited pages: pp. 10–12.

OpenAI et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

LG AI Research. 2024. Exaone 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:https://arxiv.org/abs/2412.04862*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

James Romoser. 2023. No, ruth bader ginsburg did not dissent in obergefell—and other things chatgpt gets wrong about the supreme court. *SCOTUSblog (Jan. 26, 2023), https://www. scotusblog. com/2023/01/no-ruth-bader-ginsburg-did-notdissent-in-obergefell-and-other-things-chatgpt-gets-wrong-about-the-supreme-court*.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*.

Zhongxiang Sun. 2023. A short survey of viewing large language models in legal aspect. *arXiv preprint arXiv:2303.09136*.

Qwen Team. 2025. Qwen3.

Dietrich Trautmann, Natalia Ostapuk, Quentin Grail, Adrian Pol, Guglielmo Bonifazi, Shang Gao, and Martin Gajek. 2024. Measuring the groundedness of legal question-answering systems. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 176–186, Miami, FL, USA. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

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

Zhuo Zhang, Xiangjing Hu, Jingyuan Zhang, Yating Zhang, Hui Wang, Lizhen Qu, and Zenglin Xu. 2023. FEDLEGAL: The first real-world federated learning benchmark for legal NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3492–3507, Toronto, Canada. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34:05, pages 9701–9708.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Data Filtering and Curation

A.1 Final Inclusion Criteria

During the human expert revision phase, each QA instance is categorized as either *Pass*, *Revise*, or *Hold*. Following this revision phase, all instances are evaluated by three legal experts according to the five evaluation criteria described in Appendix E. Instances labeled as *Pass* or *Revise* are included in the final dataset if they satisfy both of the following scoring conditions during the evaluation phase:

- **Average score threshold:** The instance has to receive an average score of at least 3.0 from the three annotators for each of the following criteria: **Fluency**, **Practicality**, **Relevance**, and **Legal Accuracy**.

- **Complexity threshold:** At least two of the three annotators have to assign a complexity as True, indicating that all provided legal texts are required to solve the question.

A subset of samples initially marked as *Hold*—specifically, those for which the evaluators indicated that the issue could be resolved by adding additional statutory provisions—is also sent to the evaluation stage with the necessary legal context attached. As with the *Pass/Revise* group, only samples that meet the above evaluation criteria are retained.

All other *Hold* cases are excluded from the final benchmark. This included instances where interpretation requires not only statutory provisions but also precedent-based reasoning to reach a sound legal conclusion or scenarios that are unrelated to the provided legal provisions.

Instances that fail to meet the evaluation criteria specified above are likewise excluded from the final release regardless of their initial label.

A.2 Validation Filtering Statistics

	1-hop	**2-hop**	**3-hop**	**4-hop**	**Total**
Step (2) After Reformulation	648	677	1353	357	3035
- A. Full Check	636 (98%)	647 (96%)	1273 (94%)	337 (94%)	2893 (95%)
- B. Partial Check	614 (95%)	164 (24%)	58 (4%)	4 (1%)	840 (28%)
Step (3) After LLM Val. (A ∩ B)	601 (93%)	160 (24%)	51 (4%)	4 (1%)	816 (27%)
*** Before Human Val.**	67 (10%)	158 (23%)	51 (4%)	4 (1%)	280 (9%)
Step (4&5) After Human Val.	55 (8%)	125 (18%)	46 (4%)	0 (0%)	226 (7%)

Table 7: Number of QA instances before and after each validation step. **Step (2)** refers to the number of QA instances after scenario-based reformulation (Section 3.2). **Step (3)** shows the number of instances that passed both the Partial Check and Full Check during the LLM validation stage (Section 3.3). To maintain a balanced distribution across reasoning levels, a subset of 1-hop and 2-hop instances is selectively excluded before the human validation phase (**Before Human Validation**). **Step (4&5)** presents the final number of instances that passed both human revision and evaluation (Section 3.3). Percentages represent the pass rate relative to the number of instances after scenario-based reformulation.

Table 7 presents the number of QA instances retained and filtered across each validation stage of the benchmark construction pipeline.

After scenario-based reformulation (Step 2), a total of 3,035 QA drafts are produced. During the LLM validation stage (Step 3), only 816 instances (27%) pass both the Partial and Full Checks.

The pass rate for the Partial Check drops sharply as the reasoning level increases—from 95% for 1-hop questions to just 1% for 4-hop—indicating that constructing valid multi-hop questions becomes substantially more difficult as reasoning depth increases. To compensate for this difficulty and ensure sufficient coverage at higher reasoning levels, we generate and validate additional 3-hop samples.

To maintain a balanced distribution across reasoning levels, we selectively exclude a portion of 1-hop and 2-hop instances prior to human validation (marked as * in the table).

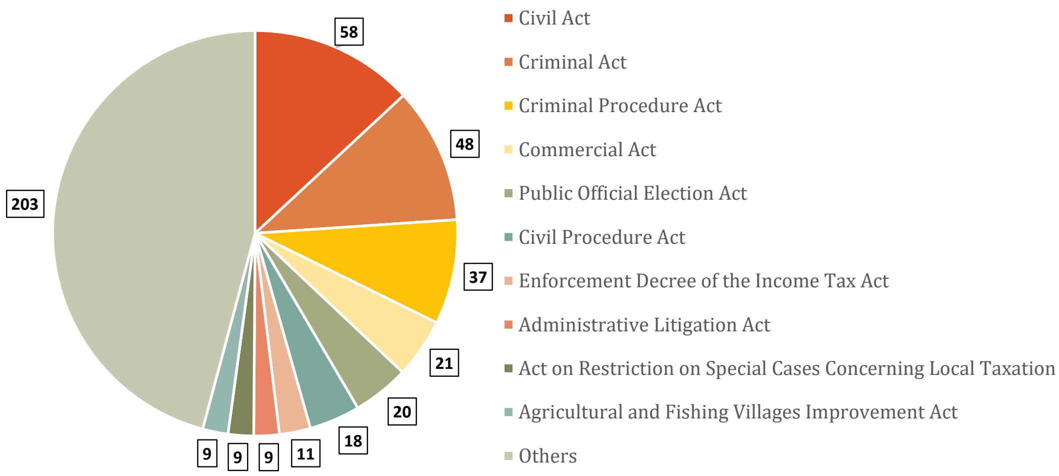

Figure 8: **Distribution of Statutes in KoBLEX.** shows the most frequently appearing statutes in the benchmark—specifically, those cited in at least nine QA instances. Statutes referenced fewer than nine times are aggregated into the *Others* category for clarity. The numbers outside each segment indicate the number of QA instances associated with each statute.

Following human expert revision and evaluation (Step 4&5), 226 QA instances remain in the final dataset. These include 55 single-hop, 125 two-hop, and 46 three-hop questions, with all 4-hop questions removed during the final filtering process described in Appendix A.1.

B List and Distribution of Statutes

Figure 8 visualizes the distribution of statutes cited in KoBLEX. For readability, only statutes that appear in at least nine QA instances are shown individually. Statutes cited fewer than nine times are aggregated into the *Others* category. The complete list of all 83 statutes and their corresponding frequencies is included in the released benchmark.

C Licensing

C.1 Source Data Licensing and Usage Rights

We utilized the public API provided by the *Korea Ministry of Government Legislation's Law Information Sharing Service*.[5] This API was used to collect both statutory provisions and precedents for the construction of our benchmark.

According to the *Act on Promotion of the Provision and Use of Public Data*, all legal information provided by the Korean Law Information Center—excluding English translations—is classified as public data. This information is openly accessible and may be freely used, including for commercial purposes, without restriction. Accordingly, our use of these sources is fully compliant with the relevant licensing and usage policies.

For the English versions of Korean statutes, we used translations provided by the Korea Legislation Research Institute (KLRI).[6] We confirmed with the KLRI that these translations may be redistributed for research purposes, provided that proper attribution is given and the usage remains non-commercial.

For statutes that are not covered by KLRI's official English translations—approximately 22 provisions—we used machine translation to generate their English versions. These machine-translated segments are explicitly marked in the dataset with the tag *%MACHINE_TRANSLATED%* to ensure transparency.

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

[6] https://elaw.klri.re.kr/eng_service/main.do

C.2 KoBLEX License

The KoBLEX benchmark is released under the **Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)** license.[7] This license permits users to copy, modify, and redistribute the dataset for non-commercial purposes as long as appropriate credit is attributed to the original authors. Any commercial use of the benchmark or its derivatives is strictly prohibited without prior written permission from the authors.

C.3 Licenses of artifacts

The EXAONE is licensed under the EXAONE AI Model License Agreement 1.1 - NC, which permits non-commercial research use only. See EXAONE-License for details. Other artifacts employed in this research are publicly available.

D Inter-Annotator Agreement Analysis

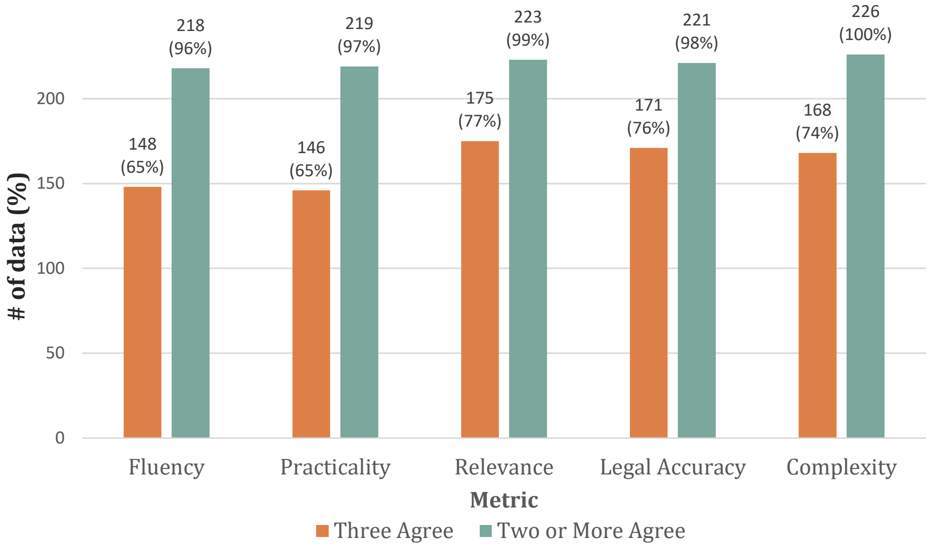

Figure 9: **Inter-annotator agreement for each evaluation metric.** We report the number and proportion of QA instances where at least two annotators agreed (*Two or More Agree*, blue) and where all three annotators provided identical labels (*Three Agree*, orange). Labels are categorized as positive (Likert scores 4–5), neutral (3), or negative (1–2).

As described in Section 3.3, three legal experts independently evaluate each revised QA instance. Fluency, practicality, relevance, and legal accuracy are rated on a 5-point Likert scale (1 to 5), while complexity is assessed as a binary label (0 or 1).

To enable agreement analysis, we re-map the Likert-scale scores into three ordinal categories, following the methodology of Fan et al. (2019):

- **Positive**: scores of 4 or 5
- **Neutral**: score of 3
- **Negative**: scores of 1 or 2

Figure 9 presents the agreement levels across the five evaluation criteria. For all metrics, over 95% of the samples exhibit agreement between at least two annotators. Moreover, more than 65% of the instances show full agreement across all three annotators, indicating a high level of reliability in the expert judgments.

[7] https://creativecommons.org/licenses/by-nc/4.0/

E Guidelines for QA Construction

This section includes detailed guidelines distributed to legal experts for revising and evaluating QA instances, as well as information on the compensation provided to annotators.

E.1 Revision Guidelines for QA Drafts

Each legal QA instance in our dataset consists of the following four components:

- **Background Scenario**: A narrative description of a specific legal situation or dispute.
- **Question**: A legal issue that naturally arises from the given scenario.
- **Answer**: A concise and logically sound response derived strictly from the provided statutes.
- **Reference Provisions**: Legal provisions that serve as the sole basis for answering the question.

Legal experts are instructed to revise each instance to ensure legal soundness and linguistic naturalness. The resulting dataset is intended to evaluate how well large language models can understand and reason in legal contexts. The revision is conducted to satisfy the following five criteria:

- **Fluency**: Sentences must be grammatically correct, coherent, and natural.
- **Practicality**: Questions should move beyond simple definitions and be framed to support practical legal reasoning.
- **Relevance**: The background and question must be closely grounded in the provided legal provisions, enabling an answer based solely on them.
- **Legal Accuracy**: Answers must faithfully and correctly interpret the given legal provisions.
- **Complexity**: Resolving the question must require using all provided legal documents without relying on external information.

Each instance is labeled with one of the following revision statuses:

- **Pass**: All components are accurate, coherent, and closely tied to the statutes. No revision is necessary.
- **Revise**: Applied when any of the following issues are identified:
  - Ungrammatical, awkward, or unnatural phrasing in any component.
  - Logical inconsistencies, such as mismatched subjects (e.g., confusion between parties) or timeline errors.
  - Redundant mention of legal article numbers or statutory content in the scenario or question.
  - Questions that are too general, factual, or fail to articulate a clear legal issue.
  - Answers that are legally incorrect, overly vague, or lack logical coherence.
  - Statutes mentioned without clearly naming the corresponding law.
- **Hold**: Used when meaningful revision is not feasible, including:
  - Extremely low-quality inputs or incoherent writing.
  - Questions that require external legal knowledge or unstated assumptions.
  - Subjective or factual questions that do not require legal interpretation.
  - Scenarios or questions that are unrelated to the provided legal provisions.

Important: All revisions must strictly adhere to the given legal statutes. Even if a different interpretation may apply in actual legal practice, annotators are instructed to provide and revise answers solely based on the scope of the provided provisions.

E.2 Evaluation Guidelines for QA Instances

This section describes the evaluation guideline used to assess the quality of the revised legal QA pairs. Evaluators are informed that the resulting dataset would be used to evaluate the legal reasoning capabilities of large language models. Each instance, previously corrected by an expert, consists of four components:

- **Background Scenario**: A narrative description of a specific legal situation or dispute.
- **Question**: A legal issue that naturally arises from the given scenario.
- **Answer**: A concise and logically sound response derived strictly from the provided statutes.
- **Reference Provisions**: Legal provisions that serve as the sole basis for answering the question.

Evaluators are instructed to assess each QA pair based on the following five criteria. Each criterion is scored independently on a 5-point Likert scale: *Excellent*, *Good*, *Fair*, *Poor*, and *Very Poor*.

The evaluation is conducted according to the following dimensions:

- **Fluency**: Sentences must be grammatically correct, coherent, and natural.
- **Practicality**: Questions should move beyond simple definitions and be framed to support practical legal reasoning.
  - E.g., generic questions like “What is the definition of a car?” should be rated low in practicality.
- **Relevance**: The background and question must be closely grounded in the provided legal provisions, enabling an answer based solely on them.
- **Legal Accuracy**: Answers must faithfully and correctly interpret the given legal provisions.
  - **Important:** This must be evaluated strictly based on the provided legal texts.
- **Complexity**: Resolving the question must require using all provided legal documents without relying on external information.
  - This is a multiple-choice evaluation item. Evaluators must select all legal provisions that are necessary to answer the question accurately. It is important to identify all reference provisions without omission.
  - If none of the provided statutes are necessary, or if additional legal provisions are needed to answer the question, evaluators should explicitly indicate this.

E.3 Compensation for Legal Expert Annotation

Each legal expert annotator is compensated with a stipend of 100,000 KRW (approximately $75 USD) for completing 80 QA instances, consisting of 20 items for revision and 60 for evaluation. Based on our internal task estimation, this amount corresponds to approximately 2.5 hours of expert work, yielding an effective hourly wage of 40,000 KRW. This rate is roughly four times the Korean legal minimum wage (as of 2025) and is deliberately set to ensure fair compensation and to attract high-quality legal annotators.

F Dataset Construction Prompt Templates

Figures 10–12 present the prompt templates used with GPT-4o during the question & answer generation and two-stage validation process. Figure 10 shows the prompts used for question & answer generation (Section 3.2); Figure 11 displays the prompt used for scenario-based reformulation (Section 3.2); and Figure 12 presents the prompts used for LLM validation (Section 3.3).

G Details of LF-EVAL

G.1 Human Evaluation Setup

To assess the effectiveness and reliability of our proposed metric LF-EVAL, we conduct a human evaluation study comparing it with various metrics including Token-F-1, BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004) and faithfulness scores (Trautmann et al., 2024).

We employ two independent groups of in-house annotators for the evaluation. All samples were anonymized so that annotators were unaware of which model produced each response. All annotators were instructed to consider:

- **Legal relevance**: Whether the response addresses the given question using the provided statutory provisions.
- **Correctness and completeness**: Whether the legal interpretation is accurate and all relevant provisions are applied appropriately.
- **Scoring**: Rate each response on a scale from 1 to 10, where
  - 1 indicates a legally irrelevant or unrelated response to the provided context.
  - 10 indicates a response that makes full and correct use of all provided statutes to reach a legally sound conclusion.

G.2 Correlation with Human Judgment

Table 8 reports Pearson correlations between automatic metrics and the averaged human scores. **Inter-annotator agreement is high** ($r_{agree} = 73.97$), indicating that the human judgments are consistent and reliable. This level of agreement supports the validity of using averaged human scores as a benchmark for evaluating automatic metrics.

Metric	**Pearson r**
LF-EVAL (Ours)	**84.90**
Faithfullness scores (Trautmann et al., 2024)	6.87
Token-level F-1	61.25
BLEU (Papineni et al., 2002)	51.89
ROUGE-L (Lin, 2004)	32.19

Table 8: Correlation of automatic metrics with human judgments. LF-EVAL achieves the strongest alignment with human evaluations.

As shown in the results, LF-EVAL exhibits a substantially stronger correlation with human judgments than other standard metrics such as Token F-1 and faithfulness scores. These findings support its effectiveness in capturing the legal fidelity of generated answers in provision-grounded QA tasks.

While the Faithfulness score has shown a strong correlation with human judgment in prior work, it performs poorly when evaluating complex multi-hop legal questions such as those in KOBLEX. This indicates that, for complex legal questions, providing detailed and strict evaluation guidelines improves the accuracy and human alignment of LLM-as-a-Judge judgments.

G.3 Example of LF-EVAL Results

Unlike conventional metrics that solely rely on gold reference answers and return only a numerical score, LF-EVAL offers both the score and an accompanying explanation that justifies the evaluation. Table 9 presents evaluation cases on KOBLEX answers generated by the GPT-4o model using PARSER, assessed through LF-EVAL. In addition to providing a fine-grained scalar score via a soft weighted sum based on G-Eval, the evaluation also reveals potential legal issues present in the generated answers.

Examples of LF-EVAL

Question: Party A and Party B, involved in a civil lawsuit over a construction contract, reached a mutual settlement recorded in the court minutes. Party A later wondered about its legal effect. What is the legal effect of a court-recorded settlement during a civil case?

Context (Gold provisions): {Civil Procedure Act Article 220 Effect of Settlement, Waiver of Claim, and Admission of Claim Article 220 (Effect of Settlement, Waiver of Claim, and Admission of Claim) When a settlement, waiver of claim, or admission of claim is recorded in the trial transcript or trial preparation date transcript, the transcript shall have the same effect as a final judgment.}

Expected output: same effect as a final judgment.

Generated Answer: If Party A and Party B reach a settlement in civil litigation in court and the details of the settlement are recorded in the pleadings, the pleadings shall **have the same effect as a final judgment.**

Score: 10, **Reason:** The actual output correctly answers the input question by stating that the record has the same effect as a final judgment, which aligns with the expected output and the legal provision in the context.

Question: Party A, who runs a small interior design business, traded 70,000 won worth of materials with Party B and wants to apply for transaction confirmation to issue a purchaser-issued invoice. What documents are required, and what conditions must be met for the transaction to be eligible?

Context (Gold provisions): {Income Tax Act Enforcement Decree Article 212-4 Issuance of Purchaser-Issued Invoices: Scope and Method, etc. Paragraph 2(2) A resident who intends to issue a purchaser-issued invoice (hereinafter referred to as "purchaser-issued invoice" in this Article) pursuant to Article 163-3(1) of the Act (hereinafter referred to as "purchaser-issued invoice" in this Article) shall issue such invoice within 30 days from the "purchaser-issued invoice") must submit an application for verification of transaction facts to the tax office director with jurisdiction over the applicant within six months from the end of the tax period in which the supply of goods or services occurred, attaching documents that objectively prove the transaction facts to the transaction fact verification application form prescribed by the Ministry of Economy and Finance. <Amended February 29, 2024>', 'Income Tax Act Enforcement Decree Article 212-4 Issuance of Purchaser-Issued Invoices: Scope and Method, etc. Paragraph 3③ Transactions subject to the transaction fact confirmation application under Paragraph 2 shall be those where the supply price per transaction is 50,000 won or more. <Amended February 29, 2024>}

Expected output: Party A must submit an application with supporting documents proving the transaction to the relevant tax office. Only transactions with a supply value of at least 50,000 won are eligible for confirmation.

Generated Answer: To issue a purchaser-issued invoice, Party A must obtain confirmation from the relevant tax office as the recipient of goods or services. Eligible transactions include cases where the supplier did not issue a tax invoice due to bankruptcy, business closure, or contract cancellation or change.

Score: 2.5344, **Reason:** The actual output correctly identifies the need for confirmation from the tax office but omits the requirement to attach objective evidence to the application. It also incorrectly states the conditions for the transaction, missing the (5만원) minimum supply value condition from the context and expected output.,

Table 9: **Examples of LF-EVAL Results.** The upper example demonstrates that the generated answer strictly follows both the provision and the expected output, justifying a perfect score of 10 with a corresponding explanation. In contrast, the lower example omits critical legal details from the expected output, which is reflected in the reasoning and results in a lower score of 2.5344.

H Experimental Details

In this section, we provide detailed descriptions of the experimental setup, including the models, evaluation metrics, and baseline methods.

H.1 Model

We utilize three LLMs in our experiments: *Qwen/Qwen3-32B* (Team, 2025), *LGAI-EXAONE/EXAONE-3.5-32B-Instruct* (Research, 2024), and *gpt-4o-2024-08-06* (OpenAI et al., 2024).

Qwen is a robust multilingual LLM trained in over 100 languages, including English and Korean, with strong reasoning capabilities. We found that Qwen3's powerful reasoning sometimes produced overly long or unfocused outputs, so we ran it in non-thinking mode to keep responses concise and on-target. EXAONE is an instruction-tuned English-Korean bilingual LLM with strong performance in Korean QA. GPT-4o is a high-performing commercial LLM demonstrating state-of-the-art capabilities across diverse tasks and languages.

To ensure reproducibility, we ran the open-source models Qwen3 and EXAONE 3.5 using the vLLM

(Kwon et al., 2023) with nucleus sampling (Finlayson et al., 2024), with temperature set to 0 and top-p to 0.9. These models were served using a single A100 GPU or two A6000 GPUs in tensor parallel mode. For GPT-4o, we used the OpenAI API with the same sampling parameters and decoding strategy, incurring a total cost of approximately $185 for all evaluations and experiments.

H.2 Evaluation Metrics

KoBLEX requires identifying the corresponding legal statutes necessary to resolve complex legal questions. Therefore, when evaluating LLM performance on KoBLEX, it is essential to assess both retrieval and generation accuracy.

Retrieval. To evaluate retrieval accuracy, we report both Exact Match (EM) and F-1 scores based on provision-level overlap between the predicted and gold statute sets. Let P be the set of predicted provisions and G the set of gold provisions. The EM score is defined as 1 if $P = G$ and 0 otherwise. To compute the F-1 score, we first calculate precision p^{prov}, recall r^{prov}, and then the F-1 score F_1^{prov}, as defined in Equation 1.

$$p^{\text{prov}} = \frac{|P \cap G|}{|P|}, \quad r^{\text{prov}} = \frac{|P \cap G|}{|G|}, \quad F_1^{\text{prov}} = \begin{cases} \frac{2p^{\text{prov}} r^{\text{prov}}}{p^{\text{prov}} + r^{\text{prov}}} & \text{if } p^{\text{prov}} + r^{\text{prov}} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Generation. To evaluate the quality of generated answers, we use both token-level F-1 and LF-Eval metrics. After normalization and word-level tokenization, the token-level F-1 score is computed by comparing the predicted and ground truth answers. Let T_p and T_g denote the sets of word-level tokens from the normalized prediction and ground truth, respectively. Precision p^{token}, recall r^{token}, and token-level F-1 score F_1^{token} are computed as follows:

$$p^{\text{token}} = \frac{|T_p \cap T_g|}{|T_p|}, \quad r^{\text{token}} = \frac{|T_p \cap T_g|}{|T_g|}, \quad F_1^{\text{token}} = \begin{cases} \frac{2p^{\text{token}} r^{\text{token}}}{p^{\text{token}} + r^{\text{token}}} & \text{if } p^{\text{token}} + r^{\text{token}} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

To compute LF-Eval, we use GPT-4o with deepeval framework (Ip and Vongthongsri, 2025) to generate a score $s \in [1, 10]$ ten times using the prompt shown in Figure 5. For each generated score s_i, we extract its associated token probability $p(s_i)$ and compute the weighted sum over the 10 generations. The final LF-Eval score is defined as:

$$\text{LF-EVAL} = \frac{1}{10} \sum_{i=1}^{10} s_i \cdot p(s_i) \tag{3}$$

H.3 Baselines

We describe the detailed implementation of our baseline methods. Existing research on retrieval-augmented reasoning has primarily focused on general-domain multi-hop QA tasks such as HotpotQA and MuSiQue. Thus, these approaches do not transfer directly to the legal-domain setting of KoBLEX. Therefore, we re-implemented each baseline by carefully analyzing the original papers and their publicly available code, adapting the logic to suit legal-domain QA better. We selected the highest-scoring examples among the filtered-out ones and refined them with the help of experts as five in-context learning demonstrations.

Retrieval Module. In our experiments, we use BM25 (Robertson and Zaragoza, 2009) retrieval over the statute corpus described in Section 6.3 across all baselines to ensure fairness in retrieval.

Standard Prompting. Standard Prompting (SP) (Brown et al., 2020) directly answers the question using only the model's parametric memory. We use standard prompting with 5-shot examples. As illustrated in Figure 13, we instruct short task descriptions followed by 5-shot examples.

Chain of Thought. Chain of Thought (CoT) (Wei et al., 2022) generates reasoning steps relying on the parametric memory before the final answer. We use Chain-of-Thought prompting with 5-shot examples. As illustrated in Figure 14, we instruct short task descriptions followed by 5-shot examples. The reasoning traces were initially generated by GPT-4o and subsequently refined by human annotators.

One-time Retrieval. One-time Retrieval (OR) uses the original question to retrieve and augment the top-n provisions, where n is the number of reasoning hops in each KoBLEX instance. We employ a BM-25 retriever for retrieving provisions. We implement retrieval module via open source toolkit (Lù, 2024). In the case of one-time retrieval, we provide the top-n provisions as retrieved-context in Figure 15, where n corresponds to the number of reasoning hops the question requires.

Self-Ask. Self-Ask (Press et al., 2023) iteratively determines whether follow-up questions are needed and generates intermediate queries. It then retrieves reference provisions and uses the augmented context to produce the final answer. We implement Self-Ask following the officially implemented code.[8] As the original implementation is designed for the general domain using the Google search engine, we adapt the retrieval module to use our own retriever instead. We set a termination condition by limiting the reasoning depth up to 5 for GPT-4o and 8 for Qwen and EXAONE. We employ the prompt in Figure 16.

IRCOT. IRCoT (Trivedi et al., 2023) interleaves CoT traces with retrieval, using the interleaved generations to incorporate external knowledge into the reasoning process. We implement IRCoT following the officially implemented code.[9] As the original implementation is designed for the general domain using the Elastic search engine, we adapt the retrieval module to use our own retriever instead. We set a termination condition by limiting the reasoning depth up to 5 for GPT-4o and 8 for Qwen and EXAONE. We employ the prompt in Figure 17.

FLARE. FLARE (Jiang et al., 2023) interleaves reasoning and retrieval similar to IRCoT but selectively performs retrieval only for reasoning steps with low confidence. We implement IRCoT following the officially implemented code.[10] As the original implementation is designed for the general domain using the Bing search engine, we adapt the retrieval module to use our own retriever instead. Due to several configurable parameters in FLARE, we set the log-probability threshold to -1.5 for Qwen and EXAONE and -0.6 for GPT-4o to achieve a retrieval rate close to 50%, which was reported as optimal in the reference paper. We adapt the instruct mode to generate retrieval queries. We set a termination condition by limiting the reasoning depth up to 5 for GPT-4o and 8 for Qwen and EXAONE. The prompt is in the Figure 17.

ProbTree. ProbTree (Cao et al., 2023) decomposes the question into a tree structure and explores multiple reasoning strategies at each node. Final answers are selected by aggregating candidates based on their generation log probabilities. We implement ProbTree following the officially implemented code.[11] As the original implementation is designed for the general domain using the Elastic search engine, we adapt the retrieval module to use our own retriever instead. For tree generation, we followed the prompt in the official implementation. We only include the retrieved context from the retrieval module when the $\{openbook\}$ is selected. Prompt used for $\{closebook, openbook, child\ aggregation\}$ is in Figure 18.

BeamAggr. BeamAggr (Chu et al., 2024) enhances ProbTree with multi-source reasoning and probabilistic answer aggregation. Since the official implementation code is unavailable, we implement BeamAggr from scratch by closely following the descriptions in the reference paper. Unlike the experiments conducted in the original paper, KoBLEX focuses on the legal domain. Accordingly, we adapt the retrieval module and restrict multi-source reasoning to closebook, openbook in order to obtain the distribution over each leaf node. Furthermore, due to the lack of sufficient sources, we utilize the log probabilities of generated tokens to estimate the response probabilities from each source. For tree generation, we followed the prompt in the official implementation of ProbTree. We only include the retrieved context from the

[8] ofirpress/self-ask

[9] StonyBrookNLP/ircot

[10] jzbjyb/FLARE

[11] THU-KEG/ProbTree

retrieval module when the $\{openbook\}$ is selected. The prompt for $\{closebook, openbook\}$ is in Figure 18.

PARSER. For parametric provision generation, we instruct the LLM to generate list-style text and extract the results using a JSON parser. For the Retrieve, Rerank, and Selection retrieval pipeline, we utilize three different language models, one for each stage. In the Retrieve stage, our retrieval module retrieves the top 100 provisions most relevant to the generated parametric provision from the statute corpus. In the Rerank stage, we use a fine-tuned BGE reranker[12](Chen et al., 2024) to sort the top 100 retrieved provisions based on relevance. In the Selection stage, an LLM generates the ID of the most relevant provision among the top 10. This selected provision is then used as the supporting legal provision.

I Case Study

To better understand how existing baseline methods perform on open-ended and multi-hop legal questions, we conduct a case study on KOBLEX. Figure 20 and 21 show English-translated responses from various baseline methods, while Figure 22 and 23 illustrate Korean responses from various baseline methods. Most baseline methods fail to reason over multiple relevant provisions. By contrast, PARSER successfully reasons over all three reference statutes and produces a coherent, legally reliable answer.

J Additional Results

	F-1		EM		Token F-1		LF-EVAL	
	Qwen-8B	EXAONE-7.8B	Qwen-8B	EXAONE-7.8B	Qwen-8B	EXAONE-7.8B	Qwen-8B	EXAONE-7.8B
SP$^{\diamondsuit}$ (Brown et al., 2020)	-	-	-	-	25.59	15.98	35.13	44.82
CoT$^{\diamondsuit}$ (Wei et al., 2022)	-	-	-	-	22.89	18.94	31.28	38.58
SP (Brown et al., 2020) + OR$^{\heartsuit}$	21.50	21.50	7.08	7.08	27.32	16.79	37.64	43.20
CoT (Wei et al., 2022) + OR$^{\heartsuit}$	21.50	21.50	7.08	7.08	26.68	22.64	39.43	43.57
Self-Ask$^{\spadesuit}$ (Press et al., 2023)	10.03	4.57	2.21	0.88	6.19	9.12	17.63	26.99
IRCoT$^{\spadesuit}$ (Trivedi et al., 2023)	21.79	19.19	4.87	3.98	12.72	19.19	20.45	24.16
FLARE$^{\spadesuit}$ (Jiang et al., 2023)	25.40	20.38	5.31	3.10	22.72	11.84	38.84	23.35
ProbTree$^{\clubsuit}$ (Cao et al., 2023)	13.26	7.76	3.10	0.88	28.53	19.61	40.93	42.30
BeamAggr$^{\clubsuit}$ (Chu et al., 2024)	11.27	5.94	2.65	0.88	17.75	10.90	33.35	31.60
PARSER$^{\clubsuit}$ (Ours)	**34.18**	**36.78**	**11.50**	**10.62**	**34.64**	**24.40**	**49.41**	**52.39**

Table 10: Experimental results of various retrieval-augmented generation methods on KOBLEX. Columns shaded in blue measure **retrieval accuracy**, and columns shaded in yellow measure **generation accuracy**. Best results are highlighted in **bold**. We utilize Qwen3-8B (Team, 2025) and EXAONE-3.5-7.8B (Research, 2024). ($\diamondsuit$: No-retrieval, $\heartsuit$: One-time retrieval, $\spadesuit$: Iterative retrieval, $\clubsuit$: Sub-query retrieval).

Table 10 shows experimental results on KOBLEX with smaller LLMs such as Qwen3-8B (Team, 2025) and EXAONE-3.5-7.8B (Research, 2024). The results show consistent patterns with robust LLMS as shown in the main body of the paper represented in Table 2. PARSER reliably achieves the highest performance across metrics. These findings confirm that our approach is effective even with limited model capacity, highlighting its scalability and generalization.

[12] dragonkue/bge-reranker-v2-m3-ko

Part 1. Instruction for Q&A Generation (1hop):

You are given a set of legal provisions (in Korean). Your task is to generate a single-hop legal question-answer pair based on the most appropriate provision from the context.

Your task is to:
1. Select one provision from the given context that is legally meaningful and suitable for generating a question.
2. Create a clear and legally relevant question that reflects the core content of the selected provision.
3. Provide an accurate legal answer that can be answered solely based on the selected provision.
4. If none of the given provisions are appropriate for generating a question, output only the line: "Not applicable"

Output Format:
question: [Generated legal question]
answer: [Accurate answer based on the provision]
selected_context: [Only include the provision used for the Q&A]

Example: {Example}

<Query>
Context: {context}

Part 2. Instruction for Q&A Generation (mhop):

You are given:
An existing Q&A pair generated from one or more legal provisions.

1. The current_context, which contains the provision(s) used for the existing Q&A.
2. The remain_context, which contains additional legal provisions not yet used.
3. Your task is to expand the original question into a (k+1)-hop question that logically incorporates one new provision from remain_context.

Your task is to:
1. Select one provision from remain_context that logically connects to the existing question or answer.
2. Generate an expanded (k+1)-hop legal question that requires both current_context and the newly selected provision to be answered.
3. Provide a new answer that integrates both contexts.
4. Output the newly selected provision as selected_context.
5. If none of the remaining provisions are appropriate for expansion, output only the line: "Not applicable"

Constraints
1. The new question must logically build upon the existing question.
2. The answer must not be answerable using only one of the contexts
(neither current_context nor the selected_context alone).
3. The question must be in Korean, concise, and naturally phrased.

Output Format:
question: [Expanded k+1-hop legal question]
answer: [New answer that depends on both current_context and selected_context]
selected_context: [One newly selected sentence from remain_context]

Example: {Example}

<Query>
Question: {question}
Answer: {answer}
Current_context: {current_context}
Remain_context: {remain_context}

Figure 10: Prompt templates used for generating legal Q&A pairs from statutory text. Part 1 describes instructions for single-hop question generation based on a single legal provision, while Part 2 outlines multi-hop question generation that requires reasoning over multiple provisions. {placeholder} indicates a slot to be filled with the corresponding value.

Part 3. Instruction for Scenario-based reformulation

You are given a legal question-answer pair based on statutory interpretation.
Rewrite this QA pair into a realistic legal scenario involving fictional characters (e.g., Person A, Person B)
where the same legal logic would apply.

Do the following:
1. Create a short but concrete fact pattern (case scenario) that would require applying the same legal reasoning.
2. Rewrite the original question to match the scenario.
3. Keep the original legal answer, with minor edits if needed to match the scenario.

Output Format:
background_scenario: {BACKGROUND SCENARIO}
question: {MULTI-HOP LEGAL QUESTION}
answer: {ANSWER}

<Query>
Question: {question}
Answer: {answer}
Context: {context}

Figure 11: Prompt template for scenario-based reformulation of legal Q&A pairs. Part 3 rewrites a statutory Q&A into a realistic case scenario while preserving the underlying legal reasoning. {placeholder} indicates a slot to be filled with the corresponding value.

Part 4. Instruction for Partial Check

You are given a question describing a legal case.

Your task is to determine whether the question can be fully answered using only the given legal provision(s) in context. Please evaluate solely based on the information in the provided context and do not assume any legal knowledge beyond it.
Use Korean for the justifications.

Input:
question: {question}
context: {context}

Output:
Answerable: (Yes / No)
Justification: (Explain briefly whether the context alone contains sufficient legal rules or logic to answer the question completely.)

<Query>
question: {question}
context: {context}

Part 5. Instruction for Full Check

You are given a legal question scenario, its proposed answer, and a set of legal context provisions.

Your task is to evaluate the following:

1. Scenario Consistent: Determine whether the background_scenario + question could have been composed using only the explicit legal content given in context. In other words, assess whether the question is logically and legally consistent with the context without requiring outside legal knowledge.

2. Correct: Evaluate whether the proposed answer is legally correct based on the background_scenario, question, and the context.

3. Derivable: Assess whether the proposed answer can be logically and completely derived from the provided context alone, without requiring any unstated assumptions.

Use Korean for the justifications.
—
Input:
background_scenario: {background_scenario}
question: {question}
answer: {answer}
context: {context}
—
Output Format:
Scenario Consistent: (Yes / No)
Scenario Justification: (Can the question be composed based solely on the given legal context? Explain with reference to the content of the legal provisions.)

Correct: (Yes / No)
Explanation: (Is the proposed answer legally appropriate in light of the scenario and the legal provisions? Provide your reasoning.)

Derivable: (Yes / No)
Justification: (Can the answer be fully derived from the legal provisions alone? Explain based on the wording and logic of the articles.)

<Query>
background_scenario: {background_scenario}
question: {question}
answer: {answer}
context: {context}

Figure 12: Prompt templates for validating legal Q&A pairs. Part 4 checks answerability based on context, while Part 5 evaluates scenario consistency, legal correctness, and derivability. {placeholder} indicates a slot to be filled with the corresponding value.

Standard Prompting prompt

You are a legal assistant AI. Given a user Background and a Question, generate a concise and accurate answer. Please provide Answer only without other explanations following the examples.

<Example>
Question: Person A was sentenced to six months in prison in the first instance for assault. Both the defendant and the prosecutor appealed. Can the appellate court sentence Person A to a longer sentence than the first trial?
Answer: The appellate court cannot sentence Person A to a longer sentence than the first trial.

{Example} (Continue to 5 examples)

<Query>
Question: {background}+{question}
Answer:

Figure 13: Prompt for **Standard Prompting (SP)**. We employ 5 different examples. {placeholder} indicates a slot to be filled with the corresponding value for evaluation.

Chain-of-Thought prompt

You are a helpful assistant. Whenever you are given a legal (or similar) question, structure your response into exactly two sections:
1. "Reasoning:"
– List each logical step you took to arrive at the conclusion.
– Use numbered or bulleted items, each beginning with a brief statement of fact or rule.
– Make the chain of thought clear and concise.
2. "Answer:"
– Provide a single, direct, definitive answer to the question in one or two sentences.
– Do not include any additional reasoning here—just the conclusion.

<Example>
Question: Person A was sentenced to six months in prison in the first instance for assault. Both the defendant and the prosecutor appealed. Can the appellate court sentence Person A to a longer sentence than the first trial?
Reasoning: In the first trial, Person A was sentenced to 6 months in prison for assault, and both the defendant and the prosecutor appealed the judgment. The scope of appellate review is limited to the "reasons for appeal" (Criminal Code §372), and the judgment can only be changed within the reasons given by the defendant and the prosecutor. Criminal Code §364(2) states that "the sentence of the final judgment may be reduced, exempted, or changed, but not aggravated, when the defendant appeals." The non-commutation principle applies. Because the prosecutor did not request an "aggravated sentence" as a ground for appeal, the appellate court cannot consider a sentence that is more severe than the first sentence. Therefore, the appellate court cannot sentence Person A to a sentence that is more severe than the first sentence.
Answer: The appellate court cannot sentence Person A to a longer sentence than the first trial.

{Example} (Continue to 5 examples)

<Query>
Question: {background}+{question}
Reasoning:

Figure 14: Prompt for **Chain-of-Thought (CoT)**. We employ 5 different examples. {placeholder} indicates a slot to be filled with the corresponding value for evaluation.

Retrieval Augmented prompt

{Standard Prompting System Prompt} or {Chain-of-Thought System Prompt}

<Example>
Context: Article 368 of the Code of Criminal Procedure Article 368 (Prohibition of adverse changes) “In cases appealed by the accused and in cases appealed on behalf of the accused, the court shall not impose a sentence heavier than that of the original judgment.”
Question: Person A was sentenced to six months in prison in the first instance for assault. Both the defendant and the prosecutor appealed.
Optional: {Reasoning: In the first trial, Person A was sentenced to 6 months in prison for assault, and both the defendant and the prosecutor appealed the judgment. The scope of appellate review is limited to the “reasons for appeal” (Criminal Code §372), and the judgment can only be changed within the reasons given by the defendant and the prosecutor. Criminal Code §364(2) states that “the sentence of the final judgment may be reduced, exempted, or changed, but not aggravated, when the defendant appeals.” The non-commutation principle applies. Because the prosecutor did not request an “aggravated sentence” as a ground for appeal, the appellate court cannot consider a sentence that is more severe than the first sentence. Therefore, the appellate court cannot sentence Person A to a sentence that is more severe than the first sentence.}
Answer: The appellate court cannot sentence Person A to a longer sentence than the first trial.

{Example} (Continue to 5 examples)

<Query>
Question: {background}+{question}
Context: {Retrieved Contexts}
Answer:

Figure 15: Prompt for retrieval augmented question answering. The system prompt and inclusion of reasoning traces vary depending on whether Self-Prediction (SP) or Chain-of-Thought (CoT) prompting is used. The retrieval method PARSER adopts the SP setting. We employ 5 different examples. {placeholder} indicates a slot to be filled with the corresponding value for evaluation.

Self-Ask prompt

You are a self-ask legal reasoning assistant. When given a new Question, follow this format exactly, with no deviations:
Question: <question>
Are follow up questions needed here: <Yes or No>
If Yes:
Follow up: <one specific clarifying question>
Intermediate answer: <brief grounded answer>
(repeat Follow up and Intermediate answer pairs until you have all facts)
So the final answer is: <your internal reasoning summary>

<Example>
Question: Person A was sentenced to six months in prison in the first instance for assault. Both the defendant and the prosecutor appealed. Can the appellate court sentence Person A to a longer sentence than the first trial?
Are follow up question needed here: Yes
Follow up: In what cases can a sentence be more severe on appeal? Intermediate answer: Because the prosecutor did not request an “aggravated sentence” as a ground for appeal, the appellate court cannot consider a sentence that is heavier than the first trial.
So the final answer is: The appellate court cannot sentence Person A to a longer sentence than the first trial.

{Example} (Continue to 5 examples)

<Query>
Question: {background}+{question}
Are follow up question needed here:

Figure 16: Prompt for **Self-ask**. We employ 5 different examples. {placeholder} indicates a slot to be filled with the corresponding value for evaluation.

IRCoT & FLARE prompt

You are a self-ask legal reasoning assistant. When given a context and a question, output exactly the following plain-text template.
Context: <legal provisions>
Question: <question>
Answer:<step-by-step reasons for the final answer> So the final answer is: <final answer>

<Example>
Context: Article 368 of the Code of Criminal Procedure Article 368 (Prohibition of adverse changes) "In cases appealed by the accused and in cases appealed on behalf of the accused, the court shall not impose a sentence heavier than that of the original judgment."
Question: Person A was sentenced to six months in prison in the first instance for assault. Both the defendant and the prosecutor appealed. Can the appellate court sentence Person A to a longer sentence than the first trial?

Answer: In the first trial, Person A was sentenced to 6 months in prison for assault, and both the defendant and the prosecutor appealed the judgment. The scope of appellate review is limited to the "reasons for appeal" (Criminal Code §372), and the judgment can only be changed within the reasons given by the defendant and the prosecutor. Criminal Code §364(2) states that "the sentence of the final judgment may be reduced, exempted, or changed, but not aggravated, when the defendant appeals." The non-commutation principle applies. Because the prosecutor did not request an "aggravated sentence" as a ground for appeal, the appellate court cannot consider a sentence that is more severe than the first sentence.
So the final answer is: The appellate court cannot sentence Person A to a sentence that is more severe than the first sentence.

{Example} (Continue to 5 examples)

<Query>
Question: {background}+{question}
Answer:

FLARE query generation prompt

The following user query has been partially masked due to low-confidence tokens.
Please review the masked query and formulate a Korean question that would allow you to search for the most relevant legal provisions needed to answer the question.

{question}
Query: {query}
New Query:

Figure 17: Prompt for **IRCoT and FLARE**. We employ 5 different examples. {placeholder} indicates a slot to be filled with the corresponding value for evaluation.

Closebook prompt

You are given legal Q&A examples. For a new legal question, answer briefly and clearly in one or two sentences.

<Example>
Question: Can the appellate court sentence Person A to a longer sentence than the first trial?
Answer: The appellate court cannot sentence Person A to a longer sentence than the first trial.

{Example} (Continue to 5 examples)

<Query>
Question: {question}
Answer:

Openbook prompt

You are given a legal question and its related law texts (context). Read the context carefully and write a concise, plain-text answer (1–2 sentences) that accurately summarizes the legal principle or outcome.

<Example>
Question: Can the appellate court sentence Person A to a longer sentence than the first trial?
Context: Article 368 of the Code of Criminal Procedure Article 368 (Prohibition of adverse changes) "In cases appealed by the accused and in cases appealed on behalf of the accused, the court shall not impose a sentence heavier than that of the original judgment."
Answer: The appellate court cannot sentence Person A to a longer sentence than the first trial.

{Example} (Continue to 5 examples)

<Query>
Question: {question}
Answer:

Child Aggregate prompt

You are given a context and a legal question.
Use only the information from the provided context to write a concise and accurate legal answer to the question.

<Example>
Context: Can the appellate court sentence Person A to a longer sentence than the first trial? The appellate court cannot sentence Person A to a longer sentence than the first trial.
Question: Person A was sentenced to six months in prison in the first instance for assault. Both the defendant and the prosecutor appealed. Can the appellate court sentence Person A to a longer sentence than the first trial?
Answer: In the first trial, Person A was sentenced to 6 months in prison for assault, and both the defendant and the prosecutor appealed the judgment. The scope of appellate review is limited to the "reasons for appeal" (Criminal Code §372), and the judgment can only be changed within the reasons given by the defendant and the prosecutor. Criminal Code §364(2) states that "the sentence of the final judgment may be reduced, exempted, or changed, but not aggravated, when the defendant appeals." The non-commutation principle applies. Because the prosecutor did not request an "aggravated sentence" as a ground for appeal, the appellate court cannot consider a sentence that is more severe than the first sentence. Therefore, the appellate court cannot sentence Person A to a sentence that is more severe than the first sentence.

{Example} (Continue to 5 examples)

<Query>
Question: {background}+{question}
Answer:

Figure 18: Prompt for **ProbTree and BeamAggr**. We employ 5 different examples. {placeholder} indicates a slot to be filled with the corresponding value for evaluation.

Parametric provision generation prompt

You are an expert legal assistant whose task is to identify and return all relevant statutory provisions that support the answer to a given legal question.
Your role is not to provide interpretations, summaries, or conclusions, but to retrieve and list the exact legal clauses that serve as a legal basis for the scenario described.
- Answer must be a list of clauses in the following format: ["Name of Law (Title or Clause Summary) Exact clause text or its key portion.","Name of Law (Title or Clause Summary) Exact clause text or its key portion.",...] without any other explanations.
- If multiple laws are involved, list all clauses together in a single list.
- If no directly applicable statutory provision exists, generate the most plausible clause in the same format, as if it were part of the relevant law.

<Example>
Question: Person A was sentenced to six months in prison in the first instance for assault. Both the defendant and the prosecutor appealed. Can the appellate court sentence Person A to a longer sentence than the first trial?
Answer: ["Article of the Code of Criminal Procedure Article (Prohibition of adverse changes) In cases appealed by the accused and in cases appealed on behalf of the accused, the court shall not impose a sentence heavier than that of the original judgment."]

{Example} (Continue to 5 examples)

<Query>
Question: {background}+{question}
Answer:

Selection prompt

You are given a question, and a list of candidate passages with associated passage IDs. Your task is to identify the most proper passage that directly support the answer to the question. Please select only one ID among given candidate passages.
Even if multiple passages seem relevant or none seem perfectly appropriate, you must select exactly one passage ID.

Background: Person A was sentenced to six months in prison in the first instance for assault. Both the defendant and the prosecutor appealed.
Question: Can the appellate court sentence Person A to a longer sentence than the first trial?
Candidates:
0: Article of the Code of Criminal Procedure Article (Prohibition of adverse changes) In cases appealed by the accused and in cases appealed on behalf of the accused, the court shall not impose a sentence heavier than that of the original judgment.
1: Article 274 of the Code of Military Justice, Cancellation of Charges, Paragraph 3 ③ Paragraphs 1 and 2 shall also apply to the withdrawal of an expression of desire for punishment in cases where the victim cannot be charged against expressed will.
...
(Top-10 provisions are placed here)
Answer: 0

{Example} (Continue to 5 examples)

<Query>
Background: {background}
Question: {question}
Context: {idx}: {provision} (Aggregate Top-10 provisions)
Answer:

Figure 19: Prompt for **ParSeR**. For parametric provision generation, we instruct LLM to generate list of parametric provisions. For selection, we instruct LLM to select most relevant provision among top-10 candidates. We employ 5 different examples. {placeholder} indicates a slot to be filled with the corresponding value for evaluation.

Background Scenario

Person A is the head of a newly established public opinion polling agency, "Trend Survey," and plans to conduct a public opinion poll related to the upcoming local elections and subsequently publish the results in the media. To this end, Person A has equipped the agency with a survey system and analysis specialists, and has fulfilled the requirements set forth by the Central Election Management Committee's regulations. Person A has submitted a written application for registration with the competent Election Survey Deliberation Commission. Person A is aware that once the registration process is completed, information about the agency will be made available to the general public.

Question

What requirements and procedures must "Trend Survey" follow to conduct and publish or report a public opinion poll related to elections? After Person A submits the registration application, what procedures must the competent Election Survey Deliberation Commission follow? Additionally, how is information about "Trend Survey" handled after the issuance of the registration certificate?

Gold Answer

"Trend Survey" must meet the requirements set by the Central Election Management Committee's regulations, such as having a survey system and analysis specialists, and must apply for registration in writing with the competent Election Survey Deliberation Commission. The competent Election Survey Deliberation Commission must process the registration within 7 days from the date of receipt of the application and issue a registration certificate. Once the registration certificate is issued, the Election Survey Deliberation Commission must promptly disclose information about the agency on the Central Election Survey Deliberation Commission's website, in accordance with the regulations set by the Central Election Management Committee.

Gold supporting provisions

PUBLIC OFFICIAL ELECTION ACT / Article. 8-9 / Registration of Polling Institutions or Organizations / Paragraph. 1
Where a polling institution or organization intends to conduct public opinion polls relating to election for the purpose of publishing or reporting, it shall file an application for the registration of the polling institution or organization with the competent deliberative committee on public opinion polls relating to election in writing after it is equipped with an examination system and experts in analysis and meets other requirements prescribed by the National Election Commission Regulations.

PUBLIC OFFICIAL ELECTION ACT / Article. 8-9 / Registration of Polling Institutions or Organizations / Paragraph. 2
The competent deliberative committee on public opinion polls relating to election upon receiving an application for registration under paragraph (1) shall accept the registration and issue a certificate of registration within seven days from the date it receives such application.

PUBLIC OFFICIAL ELECTION ACT / Article. 8-9 / Registration of Polling Institutions or Organizations / Paragraph. 3
The competent deliberative committee on public opinion polls relating to election shall without delay disclose information prescribed by the National Election Commission Regulations, which is information on the polling institution or organization (hereinafter referred to as "institution conducting public opinion polls relating to election") to which it issues a certificate of registration pursuant to paragraph (2), on the website of the National Election Survey Deliberation Commission.

Predictions (cont.)

ParSeR (Ours)
If "Trend Survey" intends to conduct public opinion polls related to elections for the purpose of publication or reporting, it must be equipped with a survey system, professional analysis personnel, and meet the requirements set forth by the rules of the National Election Commission, then submit a written registration application to the competent Election Poll Deliberation Commission. After the application is submitted by "Trend Survey," the competent commission must process the registration and issue a registration certificate within seven days from the date of receipt. Once the certificate is issued, information about "Trend Survey," as specified by the rules of the National Election Commission, must be promptly disclosed on the website of the Central Election Poll Deliberation Commission.

Figure 20: Case study of a single English instance from KoBLEX. The contents of each gold supporting provision are highlighted in different colors. ParSeR successfully responded based on a combination of three provisions.

Predictions (cont.)

Standard Prompting (SP)
If "Trend Survey" intends to conduct public opinion polls related to elections for the purpose of publication or reporting, it must meet the requirements set forth by the rules of the National Election Commission and register with the competent Election Poll Deliberation Commission. After the registration application is submitted, the commission reviews whether the requirements are met and then issues a registration certificate. Once the certificate is issued, information regarding "Trend Survey" is disclosed to the general public.

Chain-of-Thought (CoT)
If "Trend Survey" intends to conduct public opinion polls related to elections for the purpose of publication or reporting, it must meet the requirements prescribed by the rules of the National Election Commission and register with the competent Election Poll Deliberation Commission. Once the registration certificate is issued, information is disclosed to the general public.

Self-Ask
Are follow up questions needed here: Yes. (Continue)

IRCoT
"Trend Survey" must be equipped with a survey system and professional analysis personnel, and after meeting the requirements specified by the rules of the National Election Commission, it must submit a written registration application to the competent Election Poll Deliberation Commission. The competent commission must process the registration and issue a registration certificate within seven days from the date of receiving the application.

FLARE
"Trend Survey" must be equipped with a survey system and professional analysis personnel, and after meeting the requirements specified by the rules of the National Election Commission, it must submit a written registration application to the competent Election Poll Deliberation Commission. After the application is submitted, the commission reviews whether the requirements are met and issues a registration certificate.

ProbTree
If "Trend Survey" intends to conduct public opinion polls related to elections for the purpose of publication or reporting, it must meet the requirements set by the rules of the National Election Commission—such as having a survey system and professional analysis personnel—and submit a written registration application to the competent Election Poll Deliberation Commission. After the application is submitted, the commission reviews the registration in accordance with relevant procedures, and once the registration certificate is issued, information about "Trend Survey" is managed accordingly.

BeamAGgr
To conduct public opinion polls related to elections for the purpose of publication or reporting, a polling agency must meet the requirements set by the rules of the National Election Commission and submit a written registration application to the competent Election Poll Deliberation Commission. The use of virtual mobile phone numbers is permitted.

Figure 21: Case study of a single English instance from KoBLEX **(Cont.)**. While both SP and CoT capture the overall reasoning flow, they fail to identify the specific contents of the provisions. Self-Ask fail to answer the question. In IRCoT, one of the gold provisions is missing. The contents of each gold supporting provision are highlighted in different colors. FLARE and ProbTree correctly refer to one provision but miss the other two. BeamAggr provides a vague explanation and includes a hallucinated answer.

Background Scenario

갑은 신생 여론조사기관 ’트렌드서베이’의 대표로, 다가오는 지방선거와 관련된 여론조사를 실시한 후 그 결과를 언론에 공표하려는 계획을 가지고 있다. 이를 위해 갑은 조사 시스템과 분석 전문 인력을 갖추고, 중앙선거관리위원회규칙에서 요구하는 요건을 충족한 뒤 관할 선거여론조사심의위원회에 등록을 서면으로 신청하였다. 등록 절차가 완료되면 해당 기관에 관한 정보가 일반 국민에게도 공개된다는 점을 인지하고 있다.

Question

’트렌드서베이’가 선거에 관한 여론조사를 공표하거나 보도 목적으로 실시하려면 어떤 요건과 절차를 거쳐야 하며, 갑의 등록 신청 이후 관할 선거여론조사심의위원회는 어떤 처리 절차를 따라야 하는가? 또한, 등록증이 교부된 이후 ’트렌드서베이’에 관한 정보는 어떻게 처리되는가?

Gold Answer

’트렌드서베이’는 조사 시스템, 분석 전문 인력 등 중앙선거관리위원회규칙으로 정한 요건을 갖추고 관할 선거여론조사심의위원회에 서면으로 등록을 신청해야 한다. 관할 선거여론조사심의위원회는 등록신청을 접수한 날부터 7일 이내에 등록을 수리하고 등록증을 교부해야 한다. 등록증이 교부되면, 선거여론조사심의위원회는 해당 기관에 관한 정보를 중앙선거관리위원회규칙으로 정하는 바에 따라 지체 없이 중앙선거여론조사심의위원회 홈페이지에 공개해야 한다.

Gold supporting provisions

공직선거법 8조 9 여론조사 기관、단체의 등록 등 1항
① 여론조사 기관、단체가 공표 또는 보도를 목적으로 선거에 관한 여론조사를 실시하려는 때에는 조사시스템, 분석전문인력, 그 밖에 중앙선거관리위원회규칙으로 정하는 요건을 갖추어 관할 선거여론조사심의위원회에 서면으로 그 등록을 신청하여야 한다.
공직선거법 8조 9 여론조사 기관、단체의 등록 등 2항
② 제1항에 따른 등록신청을 받은 관할 선거여론조사심의위원회는 그 신청을 접수한 날부터 7일 이내에 등록을 수리하고 등록증을 교부하여야 한다.
공직선거법 8조 9 여론조사 기관、단체의 등록 등 3항
③ 선거여론조사심의위원회는 제2항에 따라 등록증을 교부한 여론조사 기관、단체(이하 "선거여론조사기관"이라 한다)에 관한 정보로서 중앙선거관리위원회규칙으로 정하는 정보를 지체 없이 중앙선거여론조사심의위원회 홈페이지에 공개하여야 한다.

ParSeR (Ours)
’트렌드서베이’가 선거에 관한 여론조사를 공표하거나 보도 목적으로 실시하려면 조사 시스템, 분석 전문 인력, 중앙선거관리위원회규칙으로 정하는 요건을 갖추어 관할 선거여론조사심의위원회에 서면으로 등록을 신청해야 한다. 갑의 등록 신청 이후, 관할 선거여론조사심의위원회는 신청을 접수한 날부터 7일 이내에 등록을 수리하고 등록증을 교부해야 한다. 등록증이 교부된 이후, ’트렌드서베이’에 관한 정보는 중앙선거관리위원회규칙으로 정하는 정보를 지체 없이 중앙선거여론조사심의위원회 홈페이지에 공개해야 한다.

Standard Prompting (SP)
’트렌드서베이’가 선거에 관한 여론조사를 공표하거나 보도 목적으로 실시하려면 중앙선거관리위원회규칙에서 정한 요건을 충족하고 관할 선거여론조사심의위원회에 등록해야 하며, 등록 신청 후 위원회는 요건 충족 여부를 심사하여 등록증을 교부한다. 등록증이 교부된 이후 ’트렌드서베이’에 관한 정보는 일반 국민에게 공개된다.

Chain-of-Thought (CoT)
’트렌드서베이’가 선거에 관한 여론조사를 공표하거나 보도 목적으로 실시하려면 중앙선거관리위원회규칙에 따른 요건을 충족하고 관할 선거여론조사심의위원회에 등록해야 하며, 등록증이 교부된 이후 기관에 관한 정보는 일반 국민에게 공개된다.

Self-Ask
Are follow up questions needed here: Yes. (Continue)

IRCoT
’트렌드서베이’는 조사 시스템과 분석 전문 인력을 갖추고, 중앙선거관리위원회규칙에서 정하는 요건을 충족한 후 관할 선거여론조사심의위원회에 서면으로 등록을 신청해야 하며, 관할 선거여론조사심의위원회는 신청을 접수한 날부터 7일 이내에 등록을 수리하고 등록증을 교부해야 한다.

Figure 22: Case study of a single Korean instance from KoBLEX. The contents of each gold supporting provision are highlighted in different colors. While both SP and CoT capture the overall reasoning flow, they fail to identify the specific contents of the provisions. Self-Ask fail to answer the question. In IRCoT, one of the gold provisions is missing.

Background Scenario

갑은 신생 여론조사기관 ’트렌드서베이’의 대표로, 다가오는 지방선거와 관련된 여론조사를 실시한 후 그 결과를 언론에 공표하려는 계획을 가지고 있다. 이를 위해 갑은 조사 시스템과 분석 전문 인력을 갖추고, 중앙선거관리위원회규칙에서 요구하는 요건을 충족한 뒤 관할 선거여론조사심의위원회에 등록을 서면으로 신청하였다. 등록 절차가 완료되면 해당 기관에 관한 정보가 일반 국민에게도 공개된다는 점을 인지하고 있다.

Question

’트렌드서베이’가 선거에 관한 여론조사를 공표하거나 보도 목적으로 실시하려면 어떤 요건과 절차를 거쳐야 하며, 갑의 등록 신청 이후 관할 선거여론조사심의위원회는 어떤 처리 절차를 따라야 하는가? 또한, 등록증이 교부된 이후 ’트렌드서베이’에 관한 정보는 어떻게 처리되는가?

Gold Answer

’트렌드서베이’는 조사 시스템, 분석 전문 인력 등 중앙선거관리위원회규칙으로 정한 요건을 갖추고 관할 선거여론조사심의위원회에 서면으로 등록을 신청해야 한다. 관할 선거여론조사심의위원회는 등록신청을 접수한 날부터 7일 이내에 등록을 수리하고 등록증을 교부해야 한다. 등록증이 교부되면, 선거여론조사심의위원회는 해당 기관에 관한 정보를 중앙선거관리위원회규칙으로 정하는 바에 따라 지체 없이 중앙선거여론조사심의위원회 홈페이지에 공개해야 한다.

Gold supporting provisions

공직선거법 8조 9 여론조사 기관、단체의 등록 등 1항
① 여론조사 기관、단체가 공표 또는 보도를 목적으로 선거에 관한 여론조사를 실시하려는 때에는 조사시스템, 분석전문인력, 그 밖에 중앙선거관리위원회규칙으로 정하는 요건을 갖추어 관할 선거여론조사심의위원회에 서면으로 그 등록을 신청하여야 한다.
공직선거법 8조 9 여론조사 기관、단체의 등록 등 2항
② 제1항에 따른 등록신청을 받은 관할 선거여론조사심의위원회는 그 신청을 접수한 날부터 7일 이내에 등록을 수리하고 등록증을 교부하여야 한다.
공직선거법 8조 9 여론조사 기관、단체의 등록 등 3항
③ 선거여론조사심의위원회는 제2항에 따라 등록증을 교부한 여론조사 기관、단체(이하 "선거여론조사기관" 이라 한다)에 관한 정보로서 중앙선거관리위원회규칙으로 정하는 정보를 지체 없이 중앙선거여론조사심의위원회 홈페이지에 공개하여야 한다.

PARSER (Ours)
’트렌드서베이’가 선거에 관한 여론조사를 공표하거나 보도 목적으로 실시하려면 조사 시스템, 분석 전문 인력, 중앙선거관리위원회규칙으로 정하는 요건을 갖추어 관할 선거여론조사심의위원회에 서면으로 등록을 신청해야 한다. 갑의 등록 신청 이후, 관할 선거여론조사심의위원회는 신청을 접수한 날부터 7일 이내에 등록을 수리하고 등록증을 교부해야 한다. 등록증이 교부된 이후, ’트렌드서베이’에 관한 정보는 중앙선거관리위원회규칙으로 정하는 정보를 지체 없이 중앙선거여론조사심의위원회 홈페이지에 공개해야 한다.

FLARE
’트렌드서베이’는 조사 시스템과 분석 전문 인력을 갖추고, 중앙선거관리위원회규칙에서 정하는 요건을 충족한 후 관할 선거여론조사심의위원회에 서면으로 등록을 신청해야 하며, 등록 신청 후 관할 위원회는 요건 충족 여부를 검토하여 등록증을 교부한다.

ProbTree
’트렌드서베이’가 선거에 관한 여론조사를 공표하거나 보도 목적으로 실시하려면 조사시스템, 분석전문인력 등 중앙선거관리위원회규칙으로 정한 요건을 갖추고 관할 선거여론조사심의위원회에 서면으로 등록을 신청해야 한다. 등록 신청 후, 관할 선거여론조사심의위원회는 관련 절차에 따라 등록 여부를 심사하고, 등록증이 교부된 이후에는 ’트렌드서베이’에 관한 정보가 관리된다.

BeamAggr
선거에 관한 여론조사를 공표하거나 보도 목적으로 실시하려면, 여론조사 기관은 중앙선거관리위원회규칙으로 정한 요건을 갖추고 관할 선거여론조사심의위원회에 서면으로 등록을 신청해야 하며, 휴대전화 가상번호를 사용할 수 있습니다.

Figure 23: Case study of a single Korean instance from KOBLEX. The contents of each gold supporting provision are highlighted in different colors. FLARE and ProbTree correctly refer to one provision but miss the other two. BeamAggr provides a vague explanation and includes a hallucinated answer.