KoLEG: On-the-Fly Korean Legal Knowledge Editing with Continuous Retrieval

Jaehyung Seo[1], **Dahyun Jung**[1], **Jaewook Lee**[1], **Yongchan Chun**[1], **Dongjun Kim**[1]
Hwijung Ryu[2], **Donghoon Shin**[2] **and Heuiseok Lim**[1†]
[1]Department of Computer Science and Engineering, Korea University
[2]KT Corporation
{seojae777,dhaabb55,jaewook133,cyc9805,junkim100,limhseok}@korea.ac.kr
{hwijung.ryu,dong-hoon.shin}@kt.com

Abstract

Korean legal knowledge is subject to frequent temporal updates driven by societal needs and government policies. Even minor modifications to legal provisions can have significant consequences, yet continuously retraining large language models (LLMs) to incorporate such updates is resource-intensive and impractical. To address this, we propose **KoLEG**, an on-the-fly Korean legal knowledge editing framework enhanced with continuous retrieval. KoLEG employs an *Editing-Aware Learning Strategy* and a *LawEdit Retriever*, which together adaptively integrate subtle linguistic nuances and continuous legislative amendments. To support this task, we construct the *Korean Legislative Amendment Dataset*, explicitly designed for continuous legal knowledge updates with attention to both temporal dynamics and linguistic subtleties. KoLEG outperforms existing locate-then-edit and retrieval-based editing methods, demonstrating superior effectiveness in legal knowledge editing while preserving linguistic capabilities. KoLEG maintains robust performance in sequential editing, improves performance on precedent application tasks, and is qualitatively validated by legal experts. The code is available at KoLEG Github.

1 Introduction

Korean legal knowledge undergoes frequent temporal updates in response to societal demands and government policies, with statutes being enacted, amended, or repealed an average of 12.2 times each (Kim, 2022; Park, 2011). For example, between 2020 and 2022 alone, over 14,000 legislative amendments were proposed, and some statutes have been revised more than 100 times[1]. Such frequent revisions present significant challenges for large language models (LLMs), which struggle to maintain temporally stable and accurate legal knowledge through pretraining alone (Kim, 2017; Hwang et al., 2022; Lai et al., 2023).

Continuously retraining LLMs to maintain up-to-date legal knowledge is both cost-prohibitive and impractical, especially given the frequency and scale of legislative changes. As legal knowledge evolves gradually and cumulatively, ongoing retraining is not a sustainable solution. To overcome these limitations, knowledge editing has been proposed as an efficient way to selectively update or correct specific knowledge in pre-trained models without full retraining (De Cao et al., 2021). Traditional knowledge editing methods, such as ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023), are not well suited for continuous updates, as their performance tends to degrade after repeated edits. More recent approaches that combine continuous learning or retrieval augmentation, including LTE (Jiang et al., 2024) and RECIPE (Chen et al., 2024b), have improved flexibility but still struggle to capture the fine-grained and subtle distinctions characteristic of legal revisions. These challenges highlight the need for a more robust and adaptive framework for legal knowledge editing.

To address these challenges, we propose **KoLEG**, a knowledge editing framework explicitly designed for the legal domain that enables on-the-fly updates to Korean legal knowledge. KoLEG integrates two core components: an **Editing-Aware Learning Strategy** and a **LawEdit Retriever**. The Editing-Aware Learning Strategy allows the model to autonomously determine when and how to perform knowledge editing, ensuring contextual adaptability and minimizing unintended performance degradation. LawEdit Retriever leverages contrastive learning to accurately retrieve relevant legal knowledge, effectively capturing subtle updates and temporal attributes. By working together, these components enable KoLEG to reference and incorporate the latest legislative amend-

[†] Corresponding Author

[1] https://klri.re.kr/lawdata/html/index.html?p=3&v=vol+2&y=2022

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9191–9217
November 4-9, 2025

ments as needed, preserving both legal accuracy and linguistic fluency. To further support continuous legal knowledge updates, we construct the **Korean Legislative Amendments Dataset**, covering 65 Korean legal categories directly collected from the Ministry of Government Legislation[2] and Lbox Open (Hwang et al., 2022). This dataset is the first to comprehensively account for real, inherently unstructured Korean legislative amendment histories, including multi-stage cumulative revisions and mappings to legal precedents based on specific timestamps.

The key contributions of this paper are:

- We propose **KoLEG**, a sustainable and on-the-fly knowledge editing framework designed for the legal domain, enabling continuous and accurate updates to Korean legal knowledge without the need for full retraining.

- KoLEG integrates an **Editing-Aware Learning Strategy** and **LawEdit Retriever** to incorporate legislative amendments and nuanced legal revisions, leveraging model internal representations for effective knowledge updates.

- We construct the **Korean Legislative Amendment Dataset**, the first benchmark for legal knowledge editing in Korean, covering cumulative updates and temporal dynamics across 65 legal categories.

- KoLEG achieves **state-of-the-art** results on legal knowledge editing tasks, with average performance of 97.7% (KULLM3) and 95.5% (Llama3.1), surpassing baselines by up to 30 percentage points. KoLEG maintains performance above 95% in sequential editing, improves precedent-to-statute matching, and is preferred by legal experts in 96% of cases.

2 Related Work

Legal Knowledge in LLMs Recent studies have investigated the application of LLMs in the legal domain, focusing on tasks such as legal document summarization, content generation, and legal consultation (Yu et al., 2022; Lai et al., 2023; Guha et al., 2024). Legal models such as SaulLM-7B (Colombo et al., 2024) and LawLLM (Shu et al., 2024) have been pre-trained on extensive English legal knowledge and have demonstrated effectiveness across various legal tasks. In Korea, LCUBE, the first Korean legal language model, has been introduced along with the Lbox benchmark dataset (Hwang et al., 2022). More recently, KBL, a Korean legal understanding benchmark covering seven legal knowledge tasks, has been presented (Kimyeeun et al., 2024). However, little research has addressed the temporal variability of legal knowledge (Kim, 2022) or how LLMs can effectively update new legal information. To address these challenges, we propose KoLEG, a framework specialized for Korean legal knowledge editing and sequential legal amendments. Additionally, we introduce a new Korean legal knowledge editing dataset designed to capture the temporal and linguistic nuances of legislative amendments.

Knowledge Editing Knowledge editing offers an efficient way to update LLMs without full retraining (De Cao et al., 2021; Mitchell et al., 2022). ROME (Meng et al., 2022) locates and updates specific parametric knowledge using causal tracing and rank-one updates, while r-ROME (Gupta et al., 2024) improves ROME's stability for sequential editing. MEMIT (Meng et al., 2023) enables simultaneous, large-scale knowledge edits by modifying multiple key-value pairs across layers. GRACE (Hartvigsen et al., 2023) introduces a codebook network for continual model adaptation, and AlphaEdit (Fang et al., 2025) uses null-space constraints to better preserve unrelated knowledge during editing. Recent retrieval-augmented methods such as LTE (Jiang et al., 2024) and RECIPE (Chen et al., 2024b) support continuous updates, but struggle to capture the nuanced and cumulative nature of legal amendments. KoLEG addresses this by using a custom retriever and learning strategy specialized for legal edits, achieving state-of-the-art performance in the Korean legal domain.

3 Preliminary

Knowledge editing aims to update the knowledge stored in LLMs f through query-answer edit pairs $\{(q_i, a_i^*)\}_{i\in[1,N]}$. Given an edit pair (q_i, a_i^*), the post-edit model f^* is expected to map the input query q_i to the target answer a_i^*, such that $a_i^* = f^*(q_i)$. In this study, we design the input queries q as questions and prompts that require the LLM to memorize and update legal knowledge (Fei et al., 2024). These queries focus on titles of laws, specific articles, and how provisions have been amended in the context of legal question answer-

[2] https://law.go.kr/

ing and completion. For instance, an input query q_i includes "*What is the content of Article 29 of the Criminal Act?*" or "*Article 29 of the Criminal Act states*". The target answer a_i^* corresponds to the legal knowledge intended for updating, serving as the accurate response to the given query q_i. To evaluate the effectiveness of legal knowledge editing, we consider the following four dimensions:

(i) **Reliability (Edit Success)** measures the model's ability to generate the correct target answer a^* for a given query q. Here, a denotes the set of all possible responses, including both correct and incorrect answers. For a set of query-answer pairs D_e and an indicator function $\mathbb{1}$, reliability $\mathbb{R}^{\text{acc}}$ is defined as:

$$\mathbb{E}_{(q,a^*)\sim D_e}\mathbb{1}\left[\arg\max_a f^*(a|q) = a^*\right] \quad (1)$$

(ii) **Generality** measures the model's ability to generalize edited knowledge to rephrased or varied forms. It accounts for the fact that legal knowledge in texts can be expressed in diverse Korean morphological forms. For a set of paraphrased query-answer pairs D_g, the generality $\mathbb{G}^{\text{acc}}$ is defined as:

$$\mathbb{E}_{(q,a^*)\sim D_g\setminus D_e}\mathbb{1}\left[\arg\max_a f^*(a|q) = a^*\right] \quad (2)$$

(iii) **Portability** evaluates the model's ability to apply edited knowledge in reverse forms, such as using the content of legal provisions to infer their corresponding articles or laws. For a set of reversed query-answer pairs D_p, the portability $\mathbb{P}^{\text{acc}}$ is defined as:

$$\mathbb{E}_{(q,a^*)\sim D_p\setminus D_e}\mathbb{1}\left[\arg\max_a f^*(a|q) = a^*\right] \quad (3)$$

(iv) **Locality** estimates the ability of the model to preserve unrelated knowledge when performing edits. This metric ensures that changes made to the model for specific legal knowledge do not unintentionally affect predictions for queries outside the edit scope D_l. The locality $\mathbb{L}^{\text{acc}}$ is defined as:

$$\mathbb{E}_{(q,a^*)\sim D_l\setminus D_e}\mathbb{1}\left[\arg\max_a f^*(a|q) = f(a|q)\right] \quad (4)$$

4 Korean Legislative Amendments

Since no existing dataset incorporates Korean legislative amendment histories for knowledge editing, we construct a new dataset that allows for tracking legal revisions and applying knowledge editing methods. We collect legislative histories based on implementation periods using the Open API[3] provided by the **National Law Open Data** under the *Ministry of Government Legislation of South Korea*. The collected legal history is structured into three subsets: Knowledge Book, Test, and Editing-Aware Learning dataset. To construct the Test and Editing-Aware Learning dataset, we utilize GPT-4 omni (OpenAI, 2023). For further details on the construction process, refer to Appendix A and C.

Knowledge Book To define the scope of legal knowledge editing, we utilize 65 Korean legal categories from the LBox Open (Hwang et al., 2022) *Statute Classification Plus* dataset. From these, we extract 1,260 statutes that have undergone enactment, amendment, or repeal since 1970 via the National Law Open Data API. We select 799 samples with at least two updates to construct the Knowledge Book. As illustrated in Figure 1 and Table 7, the Knowledge Book records implementation periods to distinguish provision versions. It preserves pre-amendment laws to support historical cases, transitional provisions, and legal stability, ensuring scalability for continuous updates.

Test Dataset The test dataset is organized to evaluate the model's effectiveness in legal knowledge editing within the Knowledge Book. As shown in Table 7, it is designed to assess the four evaluation dimensions outlined in §3. To capture different types of knowledge updates, we divide the test dataset into 799 samples with a ***single recent update*** and 345 samples with ***sequential updates incorporating temporal cues***. The single recent update test set focuses on the model's ability to edit legal knowledge based on the latest amendment, while the sequential update test set includes multiple accumulated revisions of the same legal provision, requiring models to distinguish different implementation periods using temporal cues. For detailed information, refer to Appendix C.1.

Editing-Aware Learning Dataset To enable on-the-fly knowledge updates through retrieval-augmented editing, the model learns to modify its internal knowledge based on retrieved legal information. We collect legal provisions not included in the Knowledge Book and test dataset using the National Law Open Data API, ensuring the model learns knowledge editing without direct exposure to the test samples. As illustrated in Table 7, each input query is divided into samples designed to capture Reliability, Locality, Locality+ (with non-relevant knowledge), and Portability, resulting in

[3]https://open.law.go.kr/LSO/openApi/guideList.do

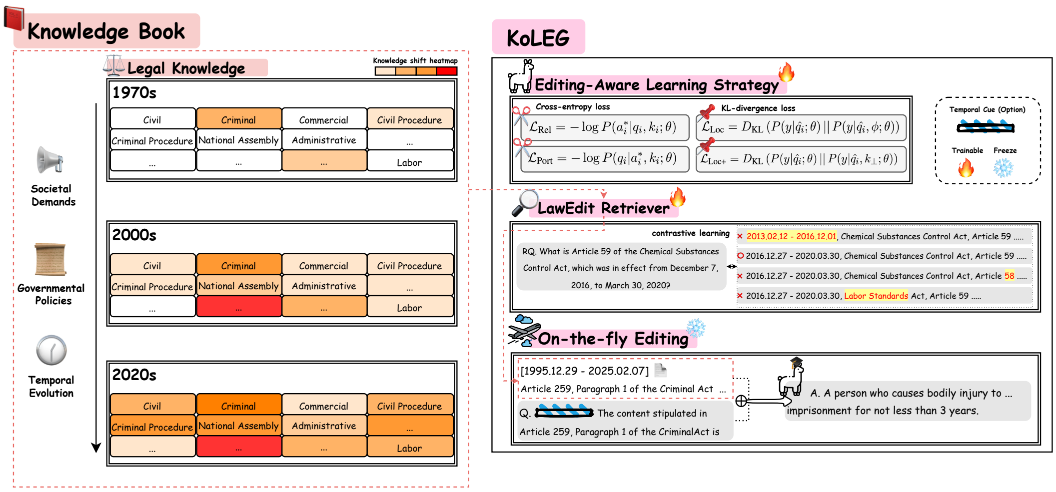

Figure 1: Overview of KoLEG, an on-the-fly Korean legal knowledge editing framework with continuous retrieval. The **Knowledge Book** stores unstructured legal provisions and amendment histories for retrieval during the implementation period. KoLEG integrates an **Editing-Aware Learning Strategy** and **LawEdit Retriever**, enabling **On-the-fly Editing** by retrieving and updating legal knowledge in real time.

a dataset of 53,392 examples. This structure enables the model to update its knowledge accordingly when retrieval-augmented information is provided while relying on its internal knowledge when the retrieved information is less relevant or absent. For detailed information, refer to Appendix C.2.

5 KoLEG

We propose KoLEG, a framework for on-the-fly updates of Korean legal knowledge that leverages continuous retrieval. KoLEG enables fine-grained knowledge editing by incorporating retrieval augmentations that account for implementation periods and subtle legal revisions. By integrating retrieved knowledge with the model's internal representations, KoLEG remains robust to continuous legislative amendments. As illustrated in Figure 1, KoLEG consists of an Editing-Aware Learning Strategy (§5.1), a LawEdit Retriever (§5.2), and an On-the-fly Editing (§5.3).

5.1 Editing-Aware Learning Strategy

KoLEG employs a retrieval-augmented learning strategy for continuous legal updates, training the model to identify and apply necessary knowledge edits. The learning strategy integrates LoRA (Low-Rank Adaptation) (Hu et al., 2022), optimizing knowledge editing as a multi-task process to minimize interference and preserve overall model performance. During training, we structure each query-answer pair $\{(q_i, a_i^*)\}_{i\in[1,|T|]}$ from the training dataset T into a batch that incorporates four loss functions to optimize knowledge editing. When the input query q_i is paired with the relevant target knowledge k_i, the model conducts knowledge editing to predict the target edited answer a_i^*. To achieve this, we apply cross-entropy loss, guiding the model to accurately predict a_i^*. Additionally, to enhance the bidirectional relationship between q_i and a_i^*, we introduce another cross-entropy loss, ensuring that when a_i^* is provided, the model can reconstruct the original query q_i. This approach allows the model to effectively utilize edited legal knowledge in various contexts while preserving connections to pre-existing legal amendments. The two loss functions are defined as follows:

$$\mathcal{L}_{\text{Rel}} = -\log P(a_i^*|q_i, k_i; \theta) \tag{5}$$

$$\mathcal{L}_{\text{Port}} = -\log P(q_i|a_i^*, k_i; \theta) \tag{6}$$

Given an input query $\hat{q}_i$ that is unrelated to knowledge editing or pertains to knowledge not stored in the Knowledge Book, the model generates a response without performing knowledge editing. Similarly, if an unrelated knowledge edit candidate $k_\perp$ is retrieved, the model is encouraged to rely on its internal parametric knowledge and provide robust answers, even when irrelevant or misleading information is presented. To ensure consistent handling of non-editing tasks and minimize unintended edits, we apply Kullback-Leibler divergence loss D_{KL}. We write $K \in \mathcal{K} \cup \{\varnothing\}$ for the knowledge condition given to the editing path. k_i denotes the relevant retrieved knowledge for q_i; $k_\perp$

an irrelevant (distractor) item; and $\phi \equiv k_\varnothing$ a null-knowledge placeholder used when the retriever returns no item (e.g., Top-1 score $< \tau = 0.78$). The edit gate remains active for all K (including ϕ), so gradients flow through the editing path even when no knowledge is injected.

The two loss functions are defined as follows:

$$\mathcal{L}_{\text{Loc}} = D_{\text{KL}}\left(P(y|\hat{q}_i;\theta)\,||\,P(y|\hat{q}_i,\phi;\theta)\right) \quad (7)$$

$$\mathcal{L}_{\text{Loc+}} = D_{\text{KL}}\left(P(y|\hat{q}_i;\theta)\,||\,P(y|\hat{q}_i,k_\perp;\theta)\right) \quad (8)$$

The total loss is computed batch-wise as follows:

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{\text{Rel}} + \lambda_2 \cdot \mathcal{L}_{\text{Port}} + \lambda_3 \cdot \mathcal{L}_{\text{Loc}} + \lambda_4 \cdot \mathcal{L}_{\text{Loc+}} \quad (9)$$

To ensure a balanced contribution of each loss term, we set the weighting factors uniformly as $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$. By jointly optimizing these losses, the model learns to balance knowledge updates with stability, avoiding unnecessary modifications while improving its ability to process sequential amendments in legal texts.

5.2 LawEdit Retriever

Legal knowledge is frequently updated in scope, subjects, and values, often with minimal changes in wording and repeated amendments across time. Given the high token-level similarity and subtle contextual shifts between versions, legal retrieval must be precise to ensure accurate editing. To this end, we train LawEdit Retriever using contrastive learning on the Editing-Aware Learning dataset, which is restructured into a training Knowledge Book. For each input query, we generate positive pairs using the exact legal provision and systematically construct hard negatives by modifying key factors that alter a statute's content, such as implementation period, law name, article number, clause, item, or provision content. As described in Figure 1, given a query specifying *'Article 59 of the Chemical Substances Control Act'* within a particular period, we create hard negatives by changing only one key factor in the positive sample. This approach enables similarity-based retrieval to accurately capture legal amendments, with the contrastive loss L_c prioritizing relevant updates and penalizing irrelevant retrievals, thereby enhancing the model's ability to distinguish fine-grained legal provisions.

$$L_c = -\log \frac{\exp(sim(q,p^+)/\tau)}{\sum_{p\in\{p^+\}\cup P^-} \exp(sim(q,p)/\tau)} \quad (10)$$

where $sim(q,p)$ represents the similarity score between the query q and document p, and τ is the temperature scaling factor. The term p^+ denotes the positive sample, and P^- represents a set of negative samples. To accurately capture the fine-grained differences in legal knowledge, we apply contrastive learning based on BGE unified retrieval function, which supports multi-granularity dense retrieval, producing representations at the word, sentence, and paragraph levels, rather than relying solely on lexical keyword overlap (Chen et al., 2024a). The experimental comparison of backbone models, including the rationale for selecting BGE as the LawEdit Retriever, is provided in Appendix B.

5.3 On-the-fly Editing

Given an input query and retrieved knowledge, KoLEG combines both sources to enable the model to edit its knowledge and generate the final output. We incorporate the implementation period into the input query to ensure that the retrieved legal knowledge targets a specific point in time for accurate knowledge editing. KoLEG accounts for legal variations over different periods, enabling precise temporal knowledge editing even for the same legal provision. By leveraging our trained LawEdit retriever, KoLEG utilizes only the Top-1 retrieved knowledge, eliminating inefficiencies caused by irrelevant retrievals. For details on the Top-N configuration of LawEdit Retriever, see Table 4. KoLEG retrieves the Top-1 knowledge only if its score exceeds a threshold (0.78). This threshold is set based on a 99% confidence interval, averaging the lower bound of the cosine similarity between the query and positive embedding with the upper bound for locality embeddings. If no retrieved knowledge surpasses the threshold, the model relies on its internal knowledge without editing.

6 Experimental Setup

We design experiments to demonstrate KoLEG's effectiveness in legal knowledge editing. Experimental details are provided in Appendix A.

Model Setting Our experiments include both the English-centric Llama-3.1-8B-Instruct (AI@Meta, 2024) and the Korean-centric KULLM3-10.7B (Kim et al., 2024), both of which are instruction-tuned models. These models have demonstrated strong capabilities in the open-source Korean NLP community and are widely used as the backbone.

# Edits	Method	Llama3.1					KULLM3				
		Reliability	Generality	Locality	Portability	Average	Reliability	Generality	Locality	Portability	Average
	Base	0.5609	0.5538	-	0.5550	0.5566	0.7090	0.7034	-	0.6113	0.6746
1	FT-M	**1.0000**	**0.9999**	0.4221	0.4191	0.7103	0.9988	0.9977	0.4866	0.4069	0.7225
	LoRA	0.9999	0.9996	0.5290	0.5209	0.7624	**0.9999**	**0.9984**	0.6863	0.5811	0.8164
	r-ROME	0.8209	0.7647	0.6516	0.5503	0.6969	0.8306	0.7751	0.8199	0.5894	0.7537
	MEMIT	0.8749	0.8200	0.5828	0.5414	0.7048	0.8830	0.8355	0.8136	0.5960	0.7820
	GRACE	0.9927	0.5538	**0.9375**	0.5550	0.7597	0.9669	0.7034	**0.9973**	0.6113	0.8197
	LTE	0.8709	0.8776	0.7224	0.6563	0.7818	0.9078	0.9065	0.7957	0.7145	0.8311
	RECIPE	0.5165	0.5155	0.6786	0.5213	0.5580	0.7069	0.7014	0.8991	0.6322	0.7349
	KoLEG	0.9734	0.9741	0.9117	**0.9616**	**0.9552**	0.9885	0.9891	0.9406	**0.9900**	**0.9770**
100	FT-M	0.7768	0.7749	0.4204	0.3694	0.5854	0.8823	0.8796	0.6348	0.5226	0.7298
	LoRA	0.6142	0.6120	0.4511	0.3816	0.5147	0.7881	0.7860	0.6539	0.5688	0.6992
	r-ROME	0.5851	0.5286	0.5905	0.4594	0.5409	0.1817	0.1810	0.1770	0.1264	0.1665
	MEMIT	0.5693	0.5550	0.5969	0.5561	0.5693	0.7223	0.7160	0.8174	0.5927	0.7121
	GRACE	0.5616	0.5542	**1.0000**	0.5560	0.6679	0.7096	0.7038	**1.0000**	0.6109	0.7561
	LTE	0.5058	0.5053	0.8302	0.4827	0.5810	0.6469	0.6452	0.8669	0.5735	0.6831
	RECIPE	0.5169	0.5158	0.6783	0.5199	0.5577	0.7081	0.6928	0.8987	0.6341	0.7334
	KoLEG	**0.9722**	**0.9729**	0.9097	**0.9604**	**0.9538**	**0.9871**	**0.9876**	0.9893	**0.9397**	**0.9759**
All	FT-M	0.6091	0.6099	0.4055	0.3272	0.4879	0.8152	0.8124	0.6450	0.5489	0.7054
	LoRA	0.5313	0.5310	0.4440	0.3613	0.4669	0.7121	0.7151	0.6619	0.5923	0.6704
	r-ROME	0.5429	0.4983	0.5943	0.4166	0.5130	0.0313	0.0313	0.0310	0.0345	0.0320
	MEMIT	0.5125	0.4876	0.5503	0.4240	0.4936	0.7081	0.6951	0.8082	0.5608	0.6931
	GRACE	0.5615	0.5542	**1.0000**	0.5552	0.6677	0.7093	0.7037	**0.9997**	0.6108	0.7559
	LTE	0.4974	0.4947	0.8235	0.4507	0.5666	0.6415	0.6399	0.8722	0.5452	0.6748
	RECIPE	0.5144	0.5115	0.6714	0.5250	0.5555	0.6997	0.6928	0.8984	0.6196	0.7276
	KoLEG	**0.9612**	**0.9621**	0.8953	**0.9544**	**0.9433**	**0.9810**	**0.9815**	0.9288	**0.9847**	**0.9690**

Table 1: Performance comparison of knowledge editing methods on Llama3.1 and KULLM3 for Korean legal updates, evaluated on 799 test samples with a single recent update. "# Edits" indicates single vs. mass editing, and "Base" refers to the pre-trained model without editing. **Bold** values denote the best performance.

Baseline Setting We compare **KoLEG** with editing baselines, including fine-tuning (**FT-M**) (Zhang et al., 2024) and Low-Rank Adaptation (**LoRA**) (Hu et al., 2022): **r-ROME** (Gupta et al., 2024) addresses the issue of model collapse in ROME (Meng et al., 2022), **MEMIT** (Meng et al., 2023) enables large-scale knowledge editing by updating multiple layers, **GRACE** (Hartvigsen et al., 2023) uses a codebook-based transformer layer for continual editing, **LTE** (Jiang et al., 2024) fine-tunes LLMs to apply retrieved edits selectively, and **RECIPE** (Chen et al., 2024b) enables lifelong knowledge editing through retrieval-augmented prompt learning. See Appendix G and H for supplementary experiments with **AlphaEdit** (Fang et al., 2025) and **GPT-4o-mini** (OpenAI, 2023).

7 Experimental Results

Editing Performance Table 1 presents the performance of editing methods on the single recent update Korean legislative amendments test set. FT-M and LoRA perform well in single edits, particularly in Reliability and Generality, but suffer significant model degradation in mass editing, especially in Locality and Portability. GRACE preserves the model's pre-existing knowledge (as seen in Locality) but remains limited in overall editing effectiveness. LTE and RECIPE, which represent recent retrieval-augmented knowledge editing approaches, aim to enable lifelong and continuous updates through external retrieval. Both struggle to capture fine-grained differences and cumulative legal amendments in Korean law, resulting in notable performance drops as edits accumulate. However, KoLEG maintains stable performance with a maximum drop of only 1.1%, while achieving the highest overall results across all scenarios. Compared to retrieval-based baselines, it delivers improvements ranging from +15% in conservative settings to nearly +40% in the most challenging cases. This demonstrates KoLEG's robustness and superior effectiveness in editing legal knowledge.

Editing Efficiency Table 2 presents the average editing and inference time for Llama3.1 and KULLM3, illustrating the efficiency of each method. Edit time refers to the average time (sec.) required to modify all test samples, while inference time represents the average time (sec.) taken for prediction after editing. FT-M, LoRA, r-ROME, and MEMIT involve direct parameter modifications, leading to longer editing times. Among them, r-

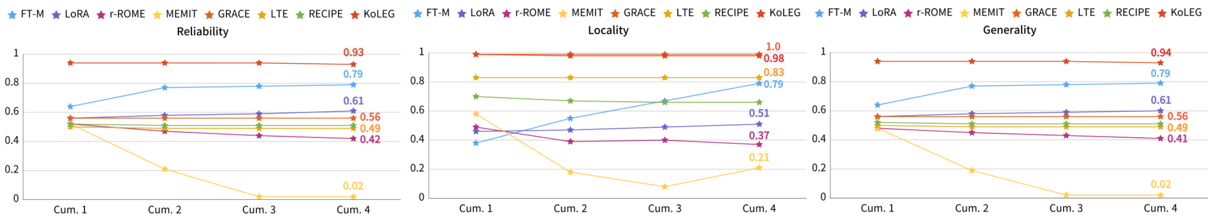

Figure 2: Performance of editing methods on 345 test samples with multiple updates and temporal cues using Llama3.1. Cum. denotes the cumulative number of updates.

Method	Edit Time	Inference Time	Total Time
FT-M	7.71	0.43	8.14
LoRA	6.46	0.44	6.90
r-ROME	62.03	0.10	62.13
MEMIT	6.98	0.06	7.04
GRACE	0.01	0.06	0.07
LTE	0.00	0.08	0.08
RECIPE	0.01	0.15	0.16
KoLEG	0.00	0.11	0.11

Table 2: Average edit and inference time (in seconds) required to update all single recent update test samples.

ROME is particularly slow, as its complex global weight modifications and lack of built-in mass editing support require additional processing. However, GRACE applies only lightweight updates to additional network layers, and LTE/RECIPE/KoLEG minimize editing time through on-the-fly updates. Regarding inference time, KoLEG exhibits a minor delay attributed to retrieval-augmented knowledge processing, similar to LTE and RECIPE. Compared to other editing methods, this difference is negligible or even advantageous in certain cases. These results indicate that KoLEG maintains efficient inference while ensuring effective on-the-fly editing.

Sequential Editing Figure 2 shows the performance of editing methods on sequential editing test samples with multiple updates and temporal cues using Llama3.1. The experiment assesses how models adapt to cumulative legal provision updates across different periods, excluding portability evaluation since predicting temporal cues is not the focus. Models generate appropriate edits at each cumulative update step (Cum. #) based on queries with temporal cues. KoLEG maintains an average performance of over 95% even after four cumulative edits, outperforming FT-M (79%) and GRACE (70%) by more than 15%. This result demonstrates that KoLEG is designed to retain stable performance even in the presence of continuous legal amendments. r-ROME and MEMIT show a sharp performance decline with consecutive updates, whereas FT-M, which updates a larger portion of model parameters, gradually adapts and improves over multiple edits. While some editing methods show significant performance fluctuations depending on the number of accumulated updates, KoLEG remains the most stable, with performance variations within only 0.7%. This result demonstrates that KoLEG effectively handles cumulative legal amendments and ensures reliable knowledge updates even in sequential editing scenarios.

Llama3.1	Reliability	Generality	Locality	Portability
KoLEG	0.9612	0.9621	0.8953	0.9544
- w/o EALS+LER	0.7635	0.7578	0.7438	0.6697
- w/o LER	0.8219	0.8184	0.7320	0.9121
- w/o EALS	0.8811	0.8752	0.8997	0.6713
- w/o Rel.	0.9069	0.9075	0.8981	0.9538
- w/o Loc.	0.9631	0.9626	0.8954	0.9545
- w/o Loc.+	0.9615	0.9609	0.9544	0.8937
- w/o Port.	0.9624	0.9627	0.9011	0.7608
KULLM3	**Reliability**	**Generality**	**Locality**	**Portability**
KoLEG	0.9810	0.9815	0.9288	0.9847
- w/o EALS+LER	0.8604	0.8523	0.8393	0.8107
- w/o LER	0.8900	0.8880	0.8209	0.9690
- w/o EALS	0.9396	0.9321	0.9361	0.8153
- w/o Rel.	0.9544	0.9504	0.9329	0.9887
- w/o Loc.	0.9825	0.9827	0.9208	0.9838
- w/o Loc.+	0.9819	0.9823	0.9303	0.9916
- w/o Port.	0.9825	0.9819	0.9279	0.8286

Table 3: Ablation study results of the Editing-Aware Learning Strategy (EALS) and LawEdit Retriever (LER) on recent update test samples.

KoLEG (Llama3.1)	Reliability	Generality	Locality	Portability	Average
+ LawEdit (1)	0.9612	0.9621	**0.8953**	0.9544	0.9433
+ LawEdit (3)	**0.9718**	**0.9720**	0.8753	**0.9585**	**0.9444**
KoLEG (KULLM3)	**Reliability**	**Generality**	**Locality**	**Portability**	**Average**
+ LawEdit (1)	0.9810	0.9815	**0.9288**	**0.9847**	**0.9690**
+ LawEdit (3)	**0.9860**	**0.9850**	0.9199	0.9815	0.9681

Table 4: KoLEG performance based on the number of retrieved knowledge (N) obtained by the LawEdit retriever in the single recent update test set.

Ablation Study We conduct an ablation study by systematically removing each loss component from KoLEG's Editing-Aware Learning Strategy (EALS) and LawEdit Retriever (LER). As shown in Table 3, removing any key component or loss leads to clear drops in reliability, generality, and portability, demonstrating that the complete strategy is essential for robust knowledge editing. Re-

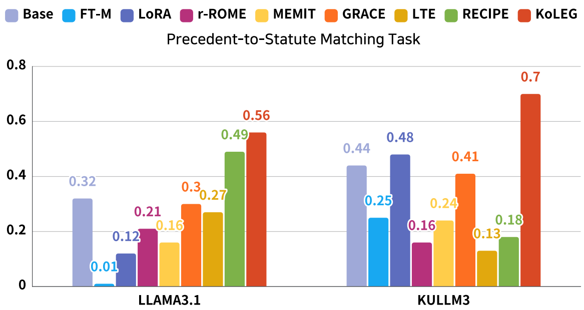

Figure 3: Evaluation of knowledge editing methods through a reasoning task predicting the correct legal provision for precedent law within a specific period. A detailed task example is provided in Appendix D.

moving the LER or both EALS and LER causes a substantial decline in locality as well as in other metrics. This is because the absence of LER leads to less precise retrieval, resulting in inaccurate or excessive editing that degrades unrelated knowledge (see Table 6). However, removing only EALS sometimes results in slightly higher locality, as the model tends to perform fewer effective edits, thus trivially preserving more non-editing knowledge. This result highlights the importance of both components: LER ensures precise and targeted editing, while EALS balances the trade-off between effective knowledge updates and minimal disruption to unrelated knowledge. The locality loss itself not only preserves the prior knowledge but also helps retain general linguistic capabilities, framing knowledge editing as a controlled mini-task that enhances real-world applicability. As shown in Figure 3 and Table 5, KoLEG maintains linguistic proficiency and performs well on legal tasks.

Top-N Retrieved Knowledge We analyze the impact of the number of Top-N retrieved knowledge items on KoLEG's performance. Table 4 shows that the average performance difference between Top-1 and Top-3 retrieval is minimal (less than 0.1%). While Top-3 retrieval slightly outperforms Top-1 for Llama3.1, KULLM3 achieves better results with Top-1. Additionally, KoLEG demonstrates that even when trained with a single retrieved knowledge, it generalizes effectively to scenarios with multiple retrieved knowledge, sometimes even improving performance. These results show that Top-1 retrieval provides an optimal balance of efficiency and effectiveness for knowledge editing, making it the default setting for KoLEG. Furthermore, depending on the specific legal knowledge or task, KoLEG can effectively leverage multiple retrieved knowledge sources while minimizing performance degradation caused by noise. In certain cases, incorporating additional retrieved knowledge even leads to performance improvements.

Precedent Application To assess the practical applicability of knowledge editing methods, we construct a precedent-to-statute matching task by collecting 500 legal precedents referencing edited statutory knowledge. Based on these precedents and their issuance dates, the task requires models to predict the citable legal provisions in a multiple-choice format. Figure 3 illustrates that most knowledge editing methods exhibit lower performance than the non-edited base model, implying that these editing methods tend to hinder the model's ability to generalize across tasks or apply edited knowledge beyond rote memorization. However, KoLEG demonstrates significant improvements, outperforming the base model by 24.2% with Llama3.1 and 25.2% with KULLM3. This result highlights KoLEG's capability not only to edit legal knowledge in real-time but also to apply edited knowledge to unseen precedents, enabling robust reasoning across broader legal domains.

Expert Evaluation To qualitatively assess knowledge editing effectiveness, we recruited three licensed Korean attorneys to blindly evaluate 50 samples each from Llama3.1 and KULLM3, edited using six different methods. As shown in Figure 11, edited outputs from LoRA, r-ROME, MEMIT, GRACE, LTE, and KoLEG were randomly shuffled, and evaluators selected the most plausible and accurate response for each sample. KoLEG was chosen as the most accurate in 97.33% of Llama3.1 cases and 94.67% of KULLM3 cases, far surpassing other methods; LTE ranked second. Other approaches frequently produced incoherent or hallucinated outputs. These results demonstrate that KoLEG provides legally valid, reliable edits while preserving generative quality better than alternatives. Detailed settings for the expert evaluation are provided in Appendix F.

Linguistic Capability Table 5 presents the evaluation results of three KoBEST (Jang et al., 2022) downstream tasks, assessing the Korean linguistic knowledge retained by the KULLM3 after editing. Both FT-M and r-ROME exhibit significant performance degradation post-editing, while LTE and LoRA show high variance across tasks, indicating

Method	KB-boolq	KB-COPA	KB-HellaSWAG	Average
Base	0.8896	0.6925	0.4612	0.6811
FT-M	0.3343	0.5277	0.2638	0.3753
LoRA	0.7025	0.6605	0.3957	0.5862
r-ROME	0.3343	0.5260	0.2470	0.3691
MEMIT	0.8577	<u>0.6943</u>	**0.4743**	0.6754
GRACE	**0.8896**	0.6925	0.4592	<u>0.6804</u>
LTE	0.3515	0.6686	0.4237	0.4813
RECIPE	0.6374	0.6935	0.4577	0.5974
KoLEG	<u>0.8688</u>	**0.7046**	<u>0.4714</u>	**0.6816**

Table 5: Korean linguistic capability evaluation of post-edited models using the KoBEST benchmark.

instability in linguistic retention. MEMIT, GRACE, and KoLEG maintain performance levels comparable to the Base model, with KoLEG achieving the highest average scores. KoLEG and MEMIT demonstrate slight performance improvements over the Base model in KB-COPA and KB-HellaSWAG. These results show that KoLEG effectively preserves linguistic capability in non-editing tasks. Furthermore, despite being designed specifically for the legal knowledge domain, KoLEG exhibits performance in a general Korean linguistic benchmark, highlighting its robustness and adaptability.

8 Conclusion

In this study, we present KoLEG, an on-the-fly Korean legal knowledge editing framework with continuous retrieval. KoLEG combines an Editing-Aware Learning Strategy and a custom LawEdit Retriever to enable precise, real-time updates of legal knowledge, even in the presence of frequent and subtle legislative amendments. By leveraging retrieval mechanisms that account for implementation periods and fine-grained legal revisions, KoLEG achieves state-of-the-art performance in the Korean legal domain while preserving linguistic capabilities. Extensive experiments demonstrate KoLEG's robustness, efficiency, and adaptability across a range of tasks. In future work, we aim to extend KoLEG to multilingual and cross-jurisdictional legal knowledge editing.

Limitation

Although KoLEG effectively updates Korean legal knowledge, there are certain limitations to consider. (i) Our research is confined to Korean legal knowledge, as legal interpretations and applications differ across jurisdictions. Legal knowledge editing is conducted within a localized context, ensuring accuracy in Korean legal amendments. In real-world applications, cross-jurisdictional legal knowledge transfer is rare, limiting the immediate applicability of KoLEG to other legal systems. Future research will extend KoLEG to multilingual settings and evaluate its effectiveness across different legal frameworks. (ii) Fully replicating baseline editing methods is challenging due to differences in language, domain, input structure, and dataset size. As a result, performance discrepancies may arise when comparing our results with English knowledge editing benchmarks, which are widely used in prior research. (iii) The scope of legal knowledge editing can be expanded further. This study does not consider case law (precedents) as "knowledge to be memorized", instead focusing on frequently revised legal provisions and their amendment histories. Since statutory law undergoes continuous changes, it serves as a more suitable target for knowledge editing than static legal definitions. (iv) The λ values in the editing-aware learning strategy and the retrieval threshold in LawEdit Retriever are adjustable hyperparameters. In this study, we adopt default settings based on design choices and statistical considerations to ensure fair evaluations. However, these parameters can be different depending on the dataset, model type, and task.

Ethics Consideration

We utilize two primary open-source legal datasets, each governed by distinct licensing policies. First, LBox Open is licensed under the Attribution-NonCommercial 4.0 International (CC BY-NC 4.0). This Creative Commons license allows the dataset to be shared and adapted for non-commercial purposes, provided proper attribution is given. Users must comply with the terms and conditions outlined in the license, which is irrevocable and grants permissions solely within the scope of copyright law. Second, National Law Open Data is provided by the *Ministry of Government Legislation of South Korea* under public data policies, ensuring open access to legal information. The dataset is freely available to all users, including for commercial purposes, without restrictions on its utilization. This aligns with the government's commitment to transparent legal information accessibility. When handling these datasets, we adhere to the respective license agreements and usage policies and ensure that all data processing and modifications comply with the original terms of use.

Acknowledgements

This work was supported by the Commercialization Promotion Agency for R&D Outcomes (COMPA) grant funded by the Korean government (Ministry of Science and ICT). This work was supported by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean government (MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-2021R1A6A1A03045425).

References

AI@Meta. 2024. Llama 3 model card.

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

Qizhou Chen, Taolin Zhang, Dongyang Li, Longtao Huang, Hui Xue, Chengyu Wang, and Xiaofeng He. 2024b. Lifelong knowledge editing for llms with retrieval-augmented continuous prompt learning. *arXiv preprint arXiv:2405.03279*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

Jingcheng Deng, Zihao Wei, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2024. Unke: Unstructured knowledge editing in large language models. *Preprint*, arXiv:2405.15349.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. Alphaedit: Null-space constrained model editing for language models. In *The Thirteenth International Conference on Learning Representations*.

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.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Akshat Gupta, Sidharth Baskaran, and Gopala Anumanchipalli. 2024. Rebuilding rome : Resolving model collapse during sequential model editing. *Preprint*, arXiv:2403.07175.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Preprint*, arXiv:2211.11031.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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.

Myeongjun Jang, Dohyung Kim, Deuk Sin Kwon, and Eric Davis. 2022. Kobest: Korean balanced evaluation of significant tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3697–3708.

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat seng Chua. 2025. Anyedit: Edit any knowledge encoded in language models. *Preprint*, arXiv:2502.05628.

Yuxin Jiang, Yufei Wang, Chuhan Wu, Wanjun Zhong, Xingshan Zeng, Jiahui Gao, Liangyou Li, Xin Jiang, Lifeng Shang, Ruiming Tang, Qun Liu, and Wei Wang. 2024. Learning to edit: Aligning LLMs with knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4689–4705, Bangkok, Thailand. Association for Computational Linguistics.

Jaeyoon Kim. 2017. Measuring the size and complexity of korean law. *Legislation and Policy Studies*, 9(1):5–29.

Jeongwook Kim, Taemin Lee, Yoonna Jang, Hyeonseok Moon, Suhyune Son, Seungyoon Lee, and Dongjun Kim. 2024. Kullm3: Korea university large language model 3. https://github.com/nlpai-lab/kullm.

Sang Yong Kim. 2022. Civil law and social changes in korea in the past 30 years. In *The Making of the Civil Codes: A Twenty-First Century Perspective*, pages 165–180. Springer.

Kimyeeun Kimyeeun, Choi Youngrok, Eunkyung Choi, JinHwan Choi, Hai 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.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. 2023. Large language models in law: A survey. *arXiv preprint arXiv:2312.03718*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. *Preprint*, arXiv:2210.07229.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast model editing at scale. In *International Conference on Learning Representations*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Soo-Chul Park. 2011. Korean law revision problems and their remedies from the perspective of the science of legislation. *Journal of Legislative Studies*, 8:27–42.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. *arXiv preprint arXiv:2407.21065*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. 2024. Easyedit: An easy-to-use knowledge editing framework for large language models. *Preprint*, arXiv:2308.07269.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.

N Zhang, Y Yao, B Tian, P Wang, S Deng, M Wang, Z Xi, S Mao, J Zhang, Y Ni, et al. 2024. A comprehensive study of knowledge editing for large language models, 2024. doi: 10.48550. *arXiv preprint arXiv.2401.01286*.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.

A Experiment Details

Implementation Details Our experiments are conducted with four NVIDIA A100 GPUs (80GB) and AMD EPYC 9334 32-core CPUs.

- **LawEdit Retriever** is trained for approximately 2 hours with a batch size of 128, where each batch consists of one query, one positive sample, and 126 hard negatives. The training is conducted using the FlagEmbedding framework[4], and retrieval is performed based on dense embeddings from the trained BGE-m3 model (Chen et al., 2024a).

- **FT-M** is implemented following the strategy of Zhang et al. (2024), where the layer identified by causal tracing is updated using cross-entropy loss applied exclusively to multi-token target answers.

- **LoRA** is applied as a low-rank adaptation to the query and value attention matrices across all layers, utilizing the AdaLoRA framework (Zhang et al., 2023).

- **r-ROME** (Gupta et al., 2024), **MEMIT** (Meng et al., 2023), and **AlphaEdit** (Fang et al., 2025) are trained using a covariance matrix C derived from Korean Wikipedia text, with a prefix-based key vector extraction strategy adapted for Korean. Causal tracing is used to determine the layers most likely to store relevant knowledge. Both methods set the covariance adjustment coefficient to 15,000 and are trained over 25 steps with a learning rate of 5×10^{-1}.

- **r-ROME** modifies a single layer, with layer 6 targeted for Llama3.1 and layer 43 for

[4] https://github.com/FlagOpen/FlagEmbedding

KULLM3. However, since this configuration degrades KULLM3's generation quality, additional experiments are conducted using the same layer settings as Llama3.1.

- **MEMIT** modifies multiple layers, applied to layers {4, 5, 6, 7, 8} in Llama 3.1 and layers {39, 40, 41, 42, 43} in KULLM3.
- **AlphaEdit** modifies layers {4, 5, 6, 7, 8} for both Llama 3.1 and KULLM3, with the null space threshold set to 2×10^{-2}.
- **UnKE** modifies layer 7 for both Llama 3.1 and KULLM3. Empirically, we find that editing a single layer is more effective than editing multiple layers.
- **AnyEdit** modifies layer 7 for both Llama 3.1 and KULLM3. The window size is set to 50 for both models.
- **GRACE** (Hartvigsen et al., 2023) is trained on the 27th MLP layer of Llama3.1 and the 43rd layer of KULLM3 for 50 steps.
- **LTE** and **RECIPE** are trained following the full-parameter fine-tuning approach described by Jiang et al. (2024) and Chen et al. (2024b), respectively.
- **KoLEG** is trained on 4 NVIDIA A100 GPUs for approximately 4 hours, optimized using AdamW with a learning rate of 3e-4, for 3 epochs with a batch size of 2.

Dataset Details As outlined in §4, we use a training set of 53,392 samples for KoLEG's Editing-Aware Learning Strategy, detailed in §5.1. For evaluation, we use 799 test samples to assess single-edit knowledge editing based on the most recent legal amendments and 345 test samples to evaluate sequential editing performance on provisions with two or more revisions. Additionally, we employ 500 test samples for the precedent-to-statute matching task and use 1,404 KB-boolq, 1,000 KB-COPA, and 500 KB-HellaSWAG test samples (Jang et al., 2022) to evaluate linguistic capability.

Evaluation Details Evaluation of LawEdit Retriever is done with MTEB (Muennighoff et al., 2022). We follow the evaluation metric proposed and utilized in MTEB. All editing evaluation is done using EasyEdit framework (Wang et al., 2024). We evaluate KoBEST (Jang et al., 2022) and the precedent-to-statute matching task using ko-lm-eval-harness[5], a framework for assessing language model performance on Korean public datasets. All evaluations are conducted in a zero-shot setting. Specifically, KB-boolq, KB-COPA, and KB-HellaSWAG are evaluated by selecting the tasks 'kobest_boolq,' 'kobest_copa,' and 'kobest_hellaswag,' respectively. The precedent-to-statute matching task follows the same evaluation setup as 'kobest_hellaswag.'

Top-1	Train	nDCG	MAP	Recall
BM25	x	0.8154	0.8154	0.8154
KoE5 (Wang et al., 2022)	x	0.7046	0.7046	0.7046
BGE-m3 (Chen et al., 2024a)	x	0.8001	0.8001	0.8001
KoE5 (Ours)	o	0.8498	0.8498	0.8498
BGE-m3 (Ours)	o	**0.9524**	**0.9524**	**0.9524**
Top-3	**Train**	**nDCG**	**MAP**	**Recall**
BM25	x	0.8826	0.8666	0.9287
KoE5 (Wang et al., 2022)	x	0.7796	0.7614	0.8329
BGE-m3 (Chen et al., 2024a)	x	0.8722	0.8551	0.9212
KoE5 (Ours)	o	0.9103	0.8964	0.9499
BGE-m3 (Ours)	o	**0.9755**	**0.9704**	**0.9899**

Table 6: Performance comparison of retrieval models for Top-1 and Top-3, evaluated on nDCG, MAP, and Recall metrics. The Train column indicates whether the model has been fine-tuned for LawEdit retriever or not.

Commercial API Details In this study, we utilized OpenAI's commercial API—specifically gpt-4o-2024-08-06 (Hurst et al., 2024)—to construct the train set, test set, sequential editing dataset, and precedent-based multiple-choice dataset among those used for our training and evaluation experiments. The cost incurred from OpenAI API calls for the GPT-4 omni model totaled \$185.67. Additionally, conducting knowledge editing via in-context learning and RAG with GPT-4o-mini resulted in a cost of \$1.84, bringing the total OpenAI API expenditure to \$187.51.

B Retriever Backbone Comparison

As described in Table 6, we compare the performance of different backbone retriever models based on their training status. Before training, dense retrievers (BGE-m3 (Chen et al., 2024a) and KoE5 (Wang et al., 2022)) perform worse than the sparse retriever (BM25). KoE5[6] achieves relatively higher performance than BGE-m3 since it is pre-trained on Korean embeddings. After training, BGE-m3 and KoE5 show an average performance increase of 48.94% and 13.64%, respectively, and outper-

[5]https://github.com/Beomi/ko-lm-evaluation-harness
[6]https://huggingface.co/nlpai-lab/KoE5

form BM25. Given that BGE-m3 achieves over 95% across all Top-1 and Top-3 metrics, we adopt it as the LawEdit Retriever backbone. Additionally, considering potential noise and inference efficiency, we set Top-1 retrieval as the default configuration.

C Korean Legislative Amendments Dataset Details

Table 7 presents the generalized structure of the Korean Legislative Amendments Dataset. For knowledge editing, KoLEG constructs a dedicated Knowledge Book, ensuring effective retrieval and updates. The test dataset is divided into two categories: single recent updates and sequential updates incorporating temporal cues to evaluate cumulative knowledge modifications. The detailed structure of the test dataset is illustrated in Figure 4 and Figure 5. The editing-aware learning dataset is structured to reflect four loss functions, with samples categorized based on the presence of relevant knowledge for training. The specific composition of this dataset is outlined in Figure 6 and Figure 7.

C.1 Test Dataset Details

Figure 4 is an example of a test set for a knowledge editing dataset based on Korean statutes. Each case in the test set consists of the original text, related questions, answers, paraphrases, restatements, and neighboring clauses for a specific, non-duplicate legal items to meet the evaluation criteria of knowledge editing. First, the CaseID is used as a unique identifier to distinguish each case, and the Info field contains the LawName, the Article, Clause, and Item, Implementation Period, and the actual Content of the provision. The Question, Answer field contains a question-answer pair that asks a content question based on the LawName and the Clause that is the target knowledge to be edited, and the Paraphrased Question contains a sentence that is semantically identical to the original question but is worded differently. The Locality field contains questions and answers about neighboring statutory provisions, and Portability contains questions and answers in a different form than those used in the Reliability evaluation, to demonstrate that the content of a particular statutory provision that has been edited can be utilized in other contexts.

Figure 5 is an example of a dataset considering sequential editing based on statutory provisions that have been amended more than once. Essentially, the components are the same as the test set,

Knowledge Book
#LawName: Act on Special Cases Concerning the Punishment of Domestic Violence Crimes
#Article: 29
#Clause: ①
#Implementation Period: 2021-01-21 — 2025-01-17
#Content: If deemed necessary for the smooth investigation and trial of domestic protection cases or for the protection of victims, a judge may issue a ruling imposing temporary measures on domestic violence offenders in accordance with any of the subparagraphs below.

Single Recent & Sequential Update Test Dataset
[Q] Reliability: [*Temporal cue (Option)*] Clause 1, Article 29 of the Act on Special ... Punishment of Violence Crimes is
[A] Reliability: If deemed ... subparagraphs below.
[Q] Generality: [*Temporal cue (Option)*] The content included in Article 29-1 of the Act ... states that
[A] Generality: If deemed ... subparagraphs below.
[Q] Locality: [*Temporal cue (Option)*] Clause 6, Article 29 of the Act on ... Punishment of Violence Crimes states that
[A] Locality: If an offender ...to protect offender.
[Q] Portability: If necessary for the efficient ... offenders as specified in the following.
[A] Portability: Clause 1, Article 29 of the Act ... Crimes.

Editing-Aware Learning Dataset (Relevant)
#Knowledge: The content of Article 33, Clause 3 of the Copyright Act, enforced from February 9, 2024, to January 17, 2025, states Hearing-impaired ... converted materials.
#Reliability Question: What does Article 33-3 of the Copyright Act state?
#Reliability Answer: Hearing-impaired individuals and their guardians, ... these converted materials.
#Portability Question: The law that includes the provision stating, "Hearing-impaired individuals ... material," is?
#Portability Answer: Copyright Act, Article 33-3.

Editing-Aware Learning Dataset (Non-Relevant)
#Knowledge (Option): The content of Article 33, Clause 3 of the Copyright Act, enforced from February 9, 2024, to January 17, 2025, states Hearing-impaired ... converted materials.
#Locality Question: What does Article 134 of the Copyright Act state?
#Locality Answer: Article 134 (Projects for Promoting a Sound Copyrighted Material Utilization Environment).

Table 7: Examples of the Korean Legislative Amendments Dataset. (Option) indicates whether a sample includes the given choice or not. *Temporal cues* are included in the test set for sequential editing. This table has been translated from Korean to English for the convenience of non-Korean speakers. (Refer to the Korean version in Figure 4, 5, 6, and 7

but the sequential editing dataset takes into account that a statutory provision is only effective for the period during which it is actually enacted, so the questions have a temporal component, such as the period of enactment or a specific point in time within the period of implementation.

C.2 Editing-Aware Learning Details

Figure 6 and Figure 7 are examples of data entered in the Editing-Aware Learning phase of KoLEG. The four factors considered in the training, Reliability, Locality, Locality+, and Portability, have in common that they contain questions and answers specific to their type. Reliability, Portability, and Locality+ are given the law information corresponding to "law_info" in the training dataset as the latest information, along with a directive to answer the given question based on the latest information provided. In contrast, Locality is given only questions and answers, without any directives or updates.

D Precedent-to-Statute Matching Task

Figure 8 presents an example of a multiple-choice dataset constructed using precedent information. To build the task, we collect precedents corresponding to the target edited legal knowledge using the National Law Open Data API[7] and Lbox Open (Hwang et al., 2022). Additionally, we extract both the incident date and ruling date from each precedent as timestamps. The incorrect answer choices are composed of other provisions within the same legal statute, ensuring a structured evaluation. This task evaluates whether models utilizing knowledge editing methods can effectively recall edited legal knowledge and infer citable legal provisions when presented with unseen precedents. Each sample contains structured entries consisting of case facts, timestamps, relevant legal provisions, legal queries, and corresponding answers.

E Prompt Examples

Figure 9 and Figure 10 are examples of the prompts we use to create our single edit dataset and sequential edit dataset, respectively. The prompts contain directives that instruct a legal professional to generate questions and answers that meet the conditions described in the prompt based on a JSON object containing the name, section number, and content of a given statute. Reliability, Generality, and Locality use the same prompt at the top, and Portability has its own example and conditions for a differently formatted question and answer than the other elements.

F Expert Evaluation Details

To conduct expert evaluation, we recruited three licensed Korean attorneys, all male in their 30s. Each evaluator was compensated at a rate of $0.70 per sample, with a total of 100 samples evaluated, resulting in a total cost of $210 across all three evaluators. Figure 11 presents examples of actual outputs generated using six different knowledge editing methods. Editing methods other than KoLEG exhibit frequent issues such as unnecessary word repetition, hallucinated content, and incoherent responses. This discrepancy between quantitatively measured editing performance and the qualitative quality of generated responses suggests that many editing methods fail to maintain output coherence. In contrast, KoLEG produces outputs that align with its measured performance, leading to an inter-annotator agreement rate exceeding 94%, with evaluators overwhelmingly selecting KoLEG as the most accurate editing method. As shown in Figure 12, for evaluation, we provided a demo interface featuring samples generated by Llama3.1 and KULLM3, accessible via a designated URL. The user guidelines for evaluators were as follows:

[7]https://open.law.go.kr/LSO/openApi/guideList.do

Llama 3.1	Reliability	Generality	Locality	Portability
AlphaEdit	0.5935	0.5658	0.8259	0.5428
KoLEG	**0.9612**	**0.9621**	**0.8953**	**0.9544**
KULLM3	**Reliability**	**Generality**	**Locality**	**Portability**
AlphaEdit	0.7338	0.6981	0.8657	0.5997
KoLEG	**0.9810**	**0.9815**	**0.9288**	**0.9847**

Table 8: Performance comparison of AlphaEdit and KoLEG on Korean legal knowledge editing for Llama 3.1 and KULLM3.

- Purpose of the Evaluation: This experiment assesses whether AI models effectively modify legal provisions without full re-training.
- Instructions: Each question pertains to a specific legal provision. The Ground Truth represents the correct legal knowledge that should be applied. Six answer choices correspond to different knowledge editing methods. Considering overall output quality, select the most accurately edited response that best aligns with the Ground Truth.

G Supplementary Results with AlphaEdit

AlphaEdit (Fang et al., 2025) is a recently proposed knowledge editing method that constrains model parameter updates to the null-space of unrelated knowledge, thereby aiming to better preserve locality during edits. Unlike MEMIT (Meng et al., 2023), which performs large-scale concurrent edits by modifying multiple key-value pairs across layers, AlphaEdit explicitly minimizes interference with non-target knowledge through null-space projection. This approach is designed to address the trade-off between edit reliability and locality, with an emphasis on robust, multi-fact editing.

Table 8 compares the performance of KoLEG and AlphaEdit on single recent legal knowledge updates for both Llama 3.1 and KULLM3. KoLEG

Llama 3.1	# Edit	Reliability	Generality	Locality
AlphaEdit	Cum. 1	0.5430	0.5391	0.7078
KoLEG	Cum. 1	**0.9397**	**0.9423**	**0.9873**
AlphaEdit	Cum. 2	0.5556	0.5690	0.7004
KoLEG	Cum. 2	**0.9404**	**0.9434**	**0.9776**
AlphaEdit	Cum. 3	0.5771	0.6021	0.6397
KoLEG	Cum. 3	**0.9367**	**0.9371**	**0.9820**
AlphaEdit	Cum. 4	0.5205	0.5714	0.7997
KoLEG	Cum. 4	**0.9350**	**0.9367**	**0.9787**
KULLM3	**# Edit**	**Reliability**	**Generality**	**Locality**
AlphaEdit	Cum. 1	0.7353	0.6932	0.8185
KoLEG	Cum. 1	**0.9642**	**0.9650**	**0.9921**
AlphaEdit	Cum. 2	0.7367	0.6831	0.8049
KoLEG	Cum. 2	**0.9642**	**0.9654**	**0.9856**
AlphaEdit	Cum. 3	0.7215	0.6770	0.8211
KoLEG	Cum. 3	**0.9617**	**0.9614**	**0.9883**
AlphaEdit	Cum. 4	0.6838	0.6643	0.8123
KoLEG	Cum. 4	**0.9606**	**0.9611**	**0.9867**

Table 9: Performance comparison of AlphaEdit and KoLEG on sequential editing tasks for Llama 3.1 and KULLM3. The higher value in each row pair (same step) is shown in bold. Cum. denotes the cumulative number of updates.

consistently outperforms AlphaEdit across all evaluation metrics. For Llama 3.1, KoLEG surpasses AlphaEdit by +37% in reliability, +40% in generalization, +7% in locality, and +41% in portability. For KULLM3, the gains are +25% in reliability, +28% in generalization, +6% in locality, and +39% in portability. These substantial improvements are observed across both models.

Table 9 shows that in sequential knowledge editing scenarios, where multiple updates are accumulated, KoLEG consistently outperforms AlphaEdit across all metrics and update steps. At every stage of sequential edits, KoLEG achieves higher reliability, generalization, and locality scores than AlphaEdit for both Llama 3.1 and KULLM3.

KoLEG's strong performance over AlphaEdit can be attributed to its domain-specialized architecture. By combining the Editing-Aware Learning Strategy and LawEdit Retriever, KoLEG is uniquely tailored to address the temporal and contextual nuances of Korean statutory amendments, challenges that are prevalent in the legal domain. While AlphaEdit excels at preserving unrelated knowledge during editing through its null-space constraints, it does not explicitly account for the fine-grained and evolving nature of legal changes, which limits its effectiveness in this context. The results demonstrate that KoLEG achieves substantially higher performance than AlphaEdit across all metrics, establishing it as a robust and domain-specialized knowledge editing method for the Korean legal domain, capable of reliable updates even under cumulative amendment scenarios.

Model	Reliability	Generality	Locality	Portability	Average
GPT-4o-mini	0.0284	0.0280	–	0.0717	0.0427
w/ BGE	0.1348	0.0661	0.4005	0.0738	0.1688
w/ LER	0.1735	0.0863	0.3780	0.0731	0.1777
w/ BGE (shot 2)	0.7769	0.0598	0.5006	0.0758	0.3533
w/ LER (shot 2)	0.8459	0.0308	0.5507	0.0724	0.3750

Table 10: Experimental results with GPT-4o-mini. "w/ BGE" denotes using BGE-M3 as the retriever, "w/ LER" denotes using LawEdit Retriever, and "shot 2" indicates the use of in-context learning with two editing examples.

H Supplementary Results with GPT-4o-mini

These supplementary experiments examine whether a powerful closed model, such as GPT-4o-mini (OpenAI, 2023), can perform on-the-fly knowledge editing simply by attaching retrieval-augmented generation (RAG) without the KoLEG framework. As shown in Table 10, GPT-4o-mini makes quantitative evaluation of knowledge editing challenging when using only RAG, as its inherently high generation diversity often leads to inconsistent or unpredictable outputs. While in-context learning with few-shot examples can improve success rates and make evaluation more reliable, the model strongly defaults to its internal parametric knowledge. Moreover, providing strict few-shot examples for every edit is impractical and limits scalability. Using a legal domain–customized retriever, such as LawEdit Retriever, instead of an off-the-shelf retriever, yields a slight improvement in retrieval and overall editing performance. The trade-off in locality observed here is consistent with trends seen in Llama3.1 and KULLM3.

Overall, applying RAG alone to closed models like GPT-4o-mini is not a practical solution for knowledge editing, as it requires the unrealistic and costly provision of strict few-shot samples for each update. In contrast, KoLEG's Editing-Aware Learning Strategy and LawEdit Retriever deliver significantly superior knowledge editing performance, even compared to augmenting GPT-4o-mini with off-the-shelf RAG. Furthermore, KoLEG can be efficiently deployed on open-source LLMs and specialized on-premise legal models, providing scalable and fluent domain-specific editing by framing knowledge editing as a mini-task.

Llama3.1	Reliability	Generality	Locality	Portability
KoLEG	0.9612	0.9621	0.8953	0.9544
- Rel. 0.5	0.9664	0.9453	0.9334	0.9453
- Loc. 0.5	0.9658	0.9455	0.9323	0.8378
- Loc.+ 0.5	0.9673	0.9462	0.9352	0.8448
- Port. 0.5	0.9659	0.9461	0.9334	0.8394

Table 11: Results of weight scheduling in the Editing-Aware Learning Strategy on recent update test samples using Llama3.1. Each row reduces the weighting factor of a single loss term to 0.5 while keeping the others at 1.0, allowing fine-grained analysis of its contribution.

Llama 3.1	Reliability	Generality	Locality	Portability
UnKE	0.5360	0.5309	0.8538	0.5227
AnyEdit	0.8149	0.7298	**0.9639**	0.4787
KoLEG	**0.9612**	**0.9621**	0.8953	**0.9544**
KULLM3	**Reliability**	**Generality**	**Locality**	**Portability**
UnKE	0.4865	0.5303	0.7256	0.4618
AnyEdit	0.5093	0.5352	0.7413	0.4634
KoLEG	**0.9810**	**0.9815**	**0.9288**	**0.9847**

Table 12: Performance comparison of unstructured editing methods and KoLEG on Korean legal knowledge editing for Llama 3.1 and KULLM3.

I Weight Scheduling for Editing-Aware Learning Strategy

We conducted additional experiments by modifying the weighting factors of the four loss terms in the Editing-Aware Learning Strategy. Table 1 presents the results under the uniform setting, where all weighting factors are fixed as $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$. Table 3 corresponds to the ablation setting, where the weighting factor of each loss is set to 0, effectively disabling its contribution. Table 11 shows the results when each λ is scheduled to 0.5, thereby halving its relative importance while keeping the others at 1.0. This scheduling allows us to perform a fine-grained analysis of each loss term's effectiveness.

The results indicate that reducing the weight of the reliability loss alleviates over-commitment to surface-level target forms, thereby improving stability on non-editing queries, albeit at the cost of slight drops in paraphrase generalization and reverse reasoning. When locality-related losses ($\mathcal{L}_{\text{Loc}}$ or $\mathcal{L}_{\text{Loc+}}$) are down-weighted, the model exhibits patterns consistent with the ablation study: locality improves, but portability performance is substantially degraded. Similarly, lowering the weight of the portability loss leads to the largest deterioration in reverse consistency.

In summary, unless the specific objective is to maximize locality, it is preferable to keep the weights of $\mathcal{L}_{\text{Port}}$, $\mathcal{L}_{\text{Loc}}$, and $\mathcal{L}_{\text{Loc+}}$ close to 1.0. A modest reduction of the reliability weight (e.g., $\lambda_1 = 0.5$) can yield practical benefits by mitigating overfitting to surface forms while maintaining overall balanced performance.

J Supplementary Results with Unstructured Editing Methods

Since KoLEG operates in an unstructured setting, we compare its performance against other unstructured knowledge-editing methods. Representative approaches include UnKE (Deng et al., 2024) and AnyEdit (Jiang et al., 2025). UnKE extends MEMIT by updating not only the weights for the next token but also for all subsequent tokens. AnyEdit generalizes this idea by employing a sliding-window mechanism that updates weights for tokens within each window. Experimental results are summarized in Table 12.

Overall, UnKE exhibited weaker performance than anticipated across all models. While AnyEdit achieved the best locality on Llama 3.1, KoLEG consistently outperformed AnyEdit by more than 10% on all other evaluation metrics for both models. Although AnyEdit's multi-token updates enable partial refinement of such information, this mechanism alone proves insufficient. These findings indicate that even the most advanced unstructured editing methods struggle to capture and revise fine-grained legal knowledge effectively, thereby underscoring the necessity of the retrieval-based editing approach adopted in KoLEG.

Split	# of statutes	# of provisions	Name of statutes used
Train	64	10,026	*NATIONAL LAND PLANNING AND UTILIZATION ACT*, *MOTOR VEHICLE MANAGEMENT ACT*, *NARCOTICS CONTROL ACT*, *ACT ON THE AGGRAVATED PUNISHMENT OF SPECIFIC CRIMES*, *MOUNTAINOUS DISTRICTS MANAGEMENT ACT*, *CHILD WELFARE ACT*, *EMERGENCY MEDICAL SERVICE ACT*, *ACT ON THE MANAGEMENT AND USE OF LIVESTOCK EXCRETA*, *INFECTIOUS DISEASE CONTROL AND PREVENTION ACT*, *ATTORNEY-AT-LAW ACT*, *ACT ON REGISTRATION OF CREDIT BUSINESS AND PROTECTION OF FINANCE USERS*, *ROAD TRAFFIC ACT*, *LABOR STANDARDS ACT*, *FOREIGN EXCHANGE TRANSACTIONS ACT*, *SUBSIDY MANAGEMENT ACT*, *ELECTRONIC FINANCIAL TRANSACTIONS ACT*, *ACT ON DOOR-TO-DOOR SALES*, *WELFARE OF SENIOR CITIZENS ACT*, *FINANCIAL INVESTMENT SERVICES AND CAPITAL MARKETS ACT*, *BUILDING ACT*, *CRIMINAL ACT*, *FARMLAND ACT*, *RESIDENT REGISTRATION ACT*, *ACT ON SPECIAL CASES CONCERNING THE PUNISHMENT OF SEXUAL CRIMES*, *FRAMEWORK ACT ON THE CONSTRUCTION INDUSTRY*, *CHEMICAL SUBSTANCES CONTROL ACT*, *ACT ON SPECIAL CASES CONCERNING THE PUNISHMENT OF CHILD ABUSE CRIMES*, *HOUSING ACT*, *ACT ON REGULATION AND PUNISHMENT OF CRIMINAL PROCEEDS CONCEALMENT*, *CUSTOMS ACT*, *IMMIGRATION ACT*, *ACT ON ELECTRONIC MONITORING*, *SPECIALIZED CREDIT FINANCE BUSINESS ACT*, *COPYRIGHT ACT*, *PUBLIC HEALTH CONTROL ACT*, *OCCUPATIONAL SAFETY AND HEALTH ACT*, *ACT ON THE PROTECTION OF CHILDREN AND YOUTH AGAINST SEX OFFENSES*, *ACT ON PROBATION*, *YOUTH PROTECTION ACT*, *NATIONAL SPORTS PROMOTION ACT*, *MILITARY SERVICE ACT*, *LICENSED REAL ESTATE AGENTS ACT*, *PROTECTION OF COMMUNICATIONS SECRETS ACT*, *FOOD SANITATION ACT*, *ACT ON SPECIAL MEASURES FOR THE CONTROL OF PUBLIC HEALTH CRIMES*, *EMPLOYMENT INSURANCE ACT*, *ACT ON THE GUARANTEE OF EMPLOYEES' RETIREMENT BENEFITS*, *PERSONAL INFORMATION PROTECTION ACT*, *RAILROAD SAFETY ACT*, *FRAMEWORK ACT ON FIREFIGHTING SERVICES*, *GAME INDUSTRY PROMOTION ACT*, *ACT ON SPECIAL MEASURES FOR DESIGNATION AND MANAGEMENT OF DEVELOPMENT RESTRICTION ZONES*, *TRADEMARK ACT*, *RESERVE FORCES ACT*, *MEDICAL SERVICE ACT*, *NATIONAL HEALTH INSURANCE ACT*, *ACT ON SPECIAL CASES CONCERNING THE PUNISHMENT OF CRIMES OF DOMESTIC VIOLENCE*, *SPECIAL ACT ON PREVENTION OF INSURANCE FRAUD*, *ACT ON REAL NAME FINANCIAL TRANSACTIONS AND CONFIDENTIALITY*, *MINIMUM WAGE ACT*, *ACT ON SPECIAL CASES CONCERNING REGULATION AND PUNISHMENT OF SPECULATIVE ACTS*, *PUNISHMENT OF MINOR OFFENSES ACT*, *ACT ON THE IMPROVEMENT OF URBAN AREAS AND RESIDENTIAL ENVIRONMENTS*, *PHARMACEUTICAL AFFAIRS ACT*
Test	60	792	*NATIONAL LAND PLANNING AND UTILIZATION ACT*, *MOTOR VEHICLE MANAGEMENT ACT*, *NARCOTICS CONTROL ACT*, *ACT ON THE AGGRAVATED PUNISHMENT OF SPECIFIC CRIMES*, *MOUNTAINOUS DISTRICTS MANAGEMENT ACT*, *CHILD WELFARE ACT*, *EMERGENCY MEDICAL SERVICE ACT*, *ACT ON THE MANAGEMENT AND USE OF LIVESTOCK EXCRETA*, *INFECTIOUS DISEASE CONTROL AND PREVENTION ACT*, *ACT ON REGISTRATION OF CREDIT BUSINESS AND PROTECTION OF FINANCE USERS*, *ATTORNEY-AT-LAW ACT*, *ROAD TRAFFIC ACT*, *LABOR STANDARDS ACT*, *FOREIGN EXCHANGE TRANSACTIONS ACT*, *ELECTRONIC FINANCIAL TRANSACTIONS ACT*, *ACT ON DOOR-TO-DOOR SALES*, *BUILDING ACT*, *WELFARE OF SENIOR CITIZENS ACT*, *FINANCIAL INVESTMENT SERVICES AND CAPITAL MARKETS ACT*, *FARMLAND ACT*, *CRIMINAL ACT*, *RESIDENT REGISTRATION ACT*, *ACT ON SPECIAL CASES CONCERNING THE PUNISHMENT OF SEXUAL CRIMES*, *FRAMEWORK ACT ON THE CONSTRUCTION INDUSTRY*, *CHEMICAL SUBSTANCES CONTROL ACT*, *ACT ON SPECIAL CASES CONCERNING THE PUNISHMENT OF CHILD ABUSE CRIMES*, *HOUSING ACT*, *CUSTOMS ACT*, *OCCUPATIONAL SAFETY AND HEALTH ACT*, *SPECIALIZED CREDIT FINANCE BUSINESS ACT*, *ACT ON ELECTRONIC MONITORING*, *PUBLIC HEALTH CONTROL ACT*, *ACT ON THE PROTECTION OF CHILDREN AND YOUTH AGAINST SEX OFFENSES*, *COPYRIGHT ACT*, *IMMIGRATION ACT*, *ACT ON PROBATION*, *NATIONAL SPORTS PROMOTION ACT*, *YOUTH PROTECTION ACT*, *MILITARY SERVICE ACT*, *PHARMACEUTICAL AFFAIRS ACT*, *LICENSED REAL ESTATE AGENTS ACT*, *PROTECTION OF COMMUNICATIONS SECRETS ACT*, *FOOD SANITATION ACT*, *EMPLOYMENT INSURANCE ACT*, *ACT ON THE GUARANTEE OF EMPLOYEES' RETIREMENT BENEFITS*, *PERSONAL INFORMATION PROTECTION ACT*, *ACT ON SPECIAL MEASURES FOR DESIGNATION AND MANAGEMENT OF DEVELOPMENT RESTRICTION ZONES*, *GAME INDUSTRY PROMOTION ACT*, *FRAMEWORK ACT ON FIREFIGHTING SERVICES*, *RAILROAD SAFETY ACT*, *TRADEMARK ACT*, *RESERVE FORCES ACT*, *MEDICAL SERVICE ACT*, *ACT ON SPECIAL CASES CONCERNING THE PUNISHMENT OF CRIMES OF DOMESTIC VIOLENCE*, *ACT ON REAL NAME FINANCIAL TRANSACTIONS AND CONFIDENTIALITY*, *MINIMUM WAGE ACT*, *ACT ON SPECIAL CASES CONCERNING REGULATION AND PUNISHMENT OF SPECULATIVE ACTS*, *PUNISHMENT OF MINOR OFFENSES ACT*, *ACT ON THE IMPROVEMENT OF URBAN AREAS AND RESIDENTIAL ENVIRONMENTS*, *NATIONAL HEALTH INSURANCE ACT*

Table 13: Number of statutes, provisions, and statute names used in the Train and Test dataset.

Split	# of statutes	# of provisions	Name of statutes used
sequential_1st	58	583	*MOTOR VEHICLE MANAGEMENT ACT, NARCOTICS CONTROL ACT, ACT ON THE AGGRAVATED PUNISHMENT OF SPECIFIC CRIMES, MOUNTAINOUS DISTRICTS MANAGEMENT ACT, CHILD WELFARE ACT, EMERGENCY MEDICAL SERVICE ACT, ACT ON THE MANAGEMENT AND USE OF LIVESTOCK EXCRETA, INFECTIOUS DISEASE CONTROL AND PREVENTION ACT, ACT ON REGISTRATION OF CREDIT BUSINESS AND PROTECTION OF FINANCE USERS, ATTORNEY-AT-LAW ACT, ROAD TRAFFIC ACT, LABOR STANDARDS ACT, FOREIGN EXCHANGE TRANSACTIONS ACT, ELECTRONIC FINANCIAL TRANSACTIONS ACT, ACT ON DOOR-TO-DOOR SALES, BUILDING ACT, WELFARE OF SENIOR CITIZENS ACT, FINANCIAL INVESTMENT SERVICES AND CAPITAL MARKETS ACT, FARMLAND ACT, CRIMINAL ACT, ACT ON SPECIAL CASES CONCERNING THE PUNISHMENT OF SEXUAL CRIMES, FRAMEWORK ACT ON THE CONSTRUCTION INDUSTRY, CHEMICAL SUBSTANCES CONTROL ACT, ACT ON SPECIAL CASES CONCERNING THE PUNISHMENT OF CHILD ABUSE CRIMES, HOUSING ACT, CUSTOMS ACT, OCCUPATIONAL SAFETY AND HEALTH ACT, SPECIALIZED CREDIT FINANCE BUSINESS ACT, ACT ON ELECTRONIC MONITORING, PUBLIC HEALTH CONTROL ACT, ACT ON PROBATION, ACT ON THE PROTECTION OF CHILDREN AND YOUTH AGAINST SEX OFFENSES, COPYRIGHT ACT, IMMIGRATION ACT, NATIONAL SPORTS PROMOTION ACT, YOUTH PROTECTION ACT, MILITARY SERVICE ACT, PHARMACEUTICAL AFFAIRS ACT, LICENSED REAL ESTATE AGENTS ACT, PROTECTION OF COMMUNICATIONS SECRETS ACT, FOOD SANITATION ACT, EMPLOYMENT INSURANCE ACT, ACT ON THE GUARANTEE OF EMPLOYEES' RETIREMENT BENEFITS, PERSONAL INFORMATION PROTECTION ACT, ACT ON SPECIAL MEASURES FOR DESIGNATION AND MANAGEMENT OF DEVELOPMENT RESTRICTION ZONES, GAME INDUSTRY PROMOTION ACT, FRAMEWORK ACT ON FIREFIGHTING SERVICES, RAILROAD SAFETY ACT, TRADEMARK ACT, RESERVE FORCES ACT, MEDICAL SERVICE ACT, ACT ON SPECIAL CASES CONCERNING THE PUNISHMENT OF CRIMES OF DOMESTIC VIOLENCE, ACT ON REAL NAME FINANCIAL TRANSACTIONS AND CONFIDENTIALITY, MINIMUM WAGE ACT, ACT ON SPECIAL CASES CONCERNING REGULATION AND PUNISHMENT OF SPECULATIVE ACTS, PUNISHMENT OF MINOR OFFENSES ACT, ACT ON THE IMPROVEMENT OF URBAN AREAS AND RESIDENTIAL ENVIRONMENTS, NATIONAL HEALTH INSURANCE ACT*
sequential_2nd	49	285	*MOTOR VEHICLE MANAGEMENT ACT, NARCOTICS CONTROL ACT, ACT ON THE AGGRAVATED PUNISHMENT OF SPECIFIC CRIMES, MOUNTAINOUS DISTRICTS MANAGEMENT ACT, CHILD WELFARE ACT, EMERGENCY MEDICAL SERVICE ACT, INFECTIOUS DISEASE CONTROL AND PREVENTION ACT, ACT ON REGISTRATION OF CREDIT BUSINESS AND PROTECTION OF FINANCE USERS, ROAD TRAFFIC ACT, LABOR STANDARDS ACT, FOREIGN EXCHANGE TRANSACTIONS ACT, ELECTRONIC FINANCIAL TRANSACTIONS ACT, ACT ON DOOR-TO-DOOR SALES, BUILDING ACT, FINANCIAL INVESTMENT SERVICES AND CAPITAL MARKETS ACT, FARMLAND ACT, ACT ON SPECIAL CASES CONCERNING THE PUNISHMENT OF SEXUAL CRIMES, FRAMEWORK ACT ON THE CONSTRUCTION INDUSTRY, CHEMICAL SUBSTANCES CONTROL ACT, HOUSING ACT, CUSTOMS ACT, OCCUPATIONAL SAFETY AND HEALTH ACT, SPECIALIZED CREDIT FINANCE BUSINESS ACT, ACT ON ELECTRONIC MONITORING, PUBLIC HEALTH CONTROL ACT, ACT ON THE PROTECTION OF CHILDREN AND YOUTH AGAINST SEX OFFENSES, COPYRIGHT ACT, YOUTH PROTECTION ACT, MILITARY SERVICE ACT, PHARMACEUTICAL AFFAIRS ACT, LICENSED REAL ESTATE AGENTS ACT, PROTECTION OF COMMUNICATIONS SECRETS ACT, FOOD SANITATION ACT, EMPLOYMENT INSURANCE ACT, ACT ON THE GUARANTEE OF EMPLOYEES' RETIREMENT BENEFITS, PERSONAL INFORMATION PROTECTION ACT, ACT ON SPECIAL MEASURES FOR DESIGNATION AND MANAGEMENT OF DEVELOPMENT RESTRICTION ZONES, GAME INDUSTRY PROMOTION ACT, FRAMEWORK ACT ON FIREFIGHTING SERVICES, RAILROAD SAFETY ACT, TRADEMARK ACT, RESERVE FORCES ACT, MEDICAL SERVICE ACT, ACT ON SPECIAL CASES CONCERNING THE PUNISHMENT OF CRIMES OF DOMESTIC VIOLENCE, MINIMUM WAGE ACT, ACT ON SPECIAL CASES CONCERNING REGULATION AND PUNISHMENT OF SPECULATIVE ACTS, PUNISHMENT OF MINOR OFFENSES ACT, ACT ON THE IMPROVEMENT OF URBAN AREAS AND RESIDENTIAL ENVIRONMENTS, NATIONAL HEALTH INSURANCE ACT*
sequential_3rd	35	145	*MOTOR VEHICLE MANAGEMENT ACT, NARCOTICS CONTROL ACT, ACT ON THE AGGRAVATED PUNISHMENT OF SPECIFIC CRIMES, MOUNTAINOUS DISTRICTS MANAGEMENT ACT, CHILD WELFARE ACT, EMERGENCY MEDICAL SERVICE ACT, INFECTIOUS DISEASE CONTROL AND PREVENTION ACT, ACT ON REGISTRATION OF CREDIT BUSINESS AND PROTECTION OF FINANCE USERS, ROAD TRAFFIC ACT, FOREIGN EXCHANGE TRANSACTIONS ACT, BUILDING ACT, FINANCIAL INVESTMENT SERVICES AND CAPITAL MARKETS ACT, FARMLAND ACT, ACT ON SPECIAL CASES CONCERNING THE PUNISHMENT OF SEXUAL CRIMES, FRAMEWORK ACT ON THE CONSTRUCTION INDUSTRY, HOUSING ACT, OCCUPATIONAL SAFETY AND HEALTH ACT, SPECIALIZED CREDIT FINANCE BUSINESS ACT, PUBLIC HEALTH CONTROL ACT, ACT ON THE PROTECTION OF CHILDREN AND YOUTH AGAINST SEX OFFENSES, YOUTH PROTECTION ACT, MILITARY SERVICE ACT, LICENSED REAL ESTATE AGENTS ACT, PROTECTION OF COMMUNICATIONS SECRETS ACT, FOOD SANITATION ACT, ACT ON SPECIAL MEASURES FOR DESIGNATION AND MANAGEMENT OF DEVELOPMENT RESTRICTION ZONES, GAME INDUSTRY PROMOTION ACT, RAILROAD SAFETY ACT, TRADEMARK ACT, RESERVE FORCES ACT, MEDICAL SERVICE ACT, MINIMUM WAGE ACT, PUNISHMENT OF MINOR OFFENSES ACT, ACT ON THE IMPROVEMENT OF URBAN AREAS AND RESIDENTIAL ENVIRONMENTS, PHARMACEUTICAL AFFAIRS ACT*
sequential_4th	25	69	*MOTOR VEHICLE MANAGEMENT ACT, NARCOTICS CONTROL ACT, MOUNTAINOUS DISTRICTS MANAGEMENT ACT, CHILD WELFARE ACT, EMERGENCY MEDICAL SERVICE ACT, INFECTIOUS DISEASE CONTROL AND PREVENTION ACT, ACT ON REGISTRATION OF CREDIT BUSINESS AND PROTECTION OF FINANCE USERS, ROAD TRAFFIC ACT, BUILDING ACT, FINANCIAL INVESTMENT SERVICES AND CAPITAL MARKETS ACT, FARMLAND ACT, FRAMEWORK ACT ON THE CONSTRUCTION INDUSTRY, HOUSING ACT, SPECIALIZED CREDIT FINANCE BUSINESS ACT, OCCUPATIONAL SAFETY AND HEALTH ACT, PUBLIC HEALTH CONTROL ACT, YOUTH PROTECTION ACT, MILITARY SERVICE ACT, FOOD SANITATION ACT, ACT ON SPECIAL MEASURES FOR DESIGNATION AND MANAGEMENT OF DEVELOPMENT RESTRICTION ZONES, GAME INDUSTRY PROMOTION ACT, TRADEMARK ACT, RESERVE FORCES ACT, MEDICAL SERVICE ACT, PHARMACEUTICAL AFFAIRS ACT*

Table 14: Number of statutes, provisions, and statute names used in the sequential editing dataset.

```
“CaseID” : 14731,
“Info” : {
    “LawName”: “자본시장과 금융투자업에 관한 법률”,
                (FINANCIAL INVESTMENT SERVICES AND CAPITAL MARKETS ACT)
    “Article”: 81,
    “Clause”: 1,
    “Item”: 4,
    “Implementation Period”: “20090204-20250117”,
    “Content”: “4. 그 밖에 투자자 보호 또는 집합투자재산의 안정적 운용 등을 해할
                우려가 있는 행위로서 대통령령으로 정하는 행위”
                (4. Any other act prescribed by Presidential Decree as likely to undermine
                the protection of investors, the stable management of collective investment
                property, etc.)
    },
“Question”: “자본시장과 금융투자업에 관한 법률의 제81조 1항 4호의 내용은
             무엇입니까?”,
             (What is the content of Article 81, Paragraph 1, Subparagraph 4 of the Act on
             the Capital Market and Financial Investment Business?)
“Answer”: “‘그 밖에 투자자 보호 또는 집합투자재산의 안정적 운용 등을 해할 우려가
           있는 행위로서 대통령령으로 정하는 행위’입니다.”,
           (It refers to “Any other act prescribed by Presidential Decree as likely to
           undermine the protection of investors, the stable management of collective
           investment property, etc.”)
“Paraphrased Question”: “자본시장과 금융투자업에 관한 법률 제 81조 1항 4호에서는
                         어떤 내용을 규정하고 있나요?”,
                         (What is stipulated in Article 81, Paragraph 1, Subparagraph 4 of
                         the Act on the Capital Market and Financial Investment Business?)
“Locality”: {
    “Question”: “자본시장과 금융투자업에 관한 법률의 제 423조의 내용은
                 무엇입니까?”,
                 (What is the content of Article 423 of the Act on the Capital Market and
                 Financial Investment Business?)
    ”Answer”:. “‘금융감독위원회는 다음 각 호의 어느 하나에 해당하는 처분 또는
                조치를 하고자 하는 경우에는 청문을 실시하여야 한다.’입니다.″
                (It states that “The Financial Services Commission shall hold a hearing to
                take any of the following dispositions or measures.”)
    },
“Portability”: {
    “Question”: “'그 밖에 투자자 보호 또는 집합투자재산의 안정적 운용 등을 해할
                 우려가 있는 행위로서 대통령령으로 정하는 행위'는 어떠한 법에
                 해당하니?”,
                 (‘Any other act prescribed by Presidential Decree as likely to undermine
                 the protection of investors, the stable management of collective investment
                 property, etc.’ falls under which law?)
“Answer”: “'자본시장과 금융투자업에 관한 법률 제81조 제1항 제4호'입니다.”,
           (It falls under Article 81, Paragraph 1, Subparagraph 4 of the Act on the Capital
           Market and Financial Investment Business.)
    }
```

Figure 4: An example of a knowledge editing test set based on Korean law, designed to evaluate the performance of a knowledge editing methodology for legal knowledge. The test set includes samples for assessing Reliability, Generality (paraphrased questions), Locality, and Portability, ensuring a comprehensive evaluation of the model's ability to apply, retain, and transfer edited legal knowledge accurately.

“CaseID” : 559,
“Info” : {
“LawName”: “의료법”,
(MEDICAL SERVICE ACT)

“Article”: 82,
“Clause”: 1,
“Item”: None,
“Implementation Period”: “20080418-20100130”,
“Content”: “①안마사는 「장애인복지법」에 따른 시각장애인 중 다음 각 호의 어느 하나에 해당하는 자로서 시ㆍ도지사에게 자격인정을 받아야 한다.”
((1) A massage therapist shall be a visually-impaired person under the Act on Welfare of Persons with Disabilities, who falls under any of the following subparagraphs and who is accredited by a relevant Mayor/Do Governor:)
},
“date”: “20080418”,
“Reliability”: {
“Question”: “2008년 4월 18일부터 2010년 3월 18일까지 시행된 의료법 제82조 제1항은 무엇인가요?”,
(What was Article 82, Paragraph 1 of the Medical Act, which was in effect from April 18, 2008, to March 18, 2010?)
“Answer”: “'안마사는 「장애인복지법」에 따른 시각장애인 중 다음 각 호의 어느 하나에 해당하는 자로서 시ㆍ도지사에게 자격인정을 받아야 한다.'입니다.”.
(It states that “(1) A massage therapist shall be a visually-impaired person under the Act on Welfare of Persons with Disabilities, who falls under any of the following subparagraphs and who is accredited by a relevant Mayor/Do Governor:”)
“Paraphrase_Question”: “의료법 제82조 제1항의 2008년 4월 18일부터 2010년 3월 18일까지의 내용은 어떤 것인가요?”,
(What provisions were included in Article 82, Paragraph 1 of the Medical Act during the period from April 18, 2008, to March 18, 2010?)
},
“Locality”: {
“Question”: “2008년 4월 18일에 시행된 의료법의 제82조 제1항 제2호는 무엇인가요?”,
(As of June 15, 2009, what was the content of Article 82, Paragraph 1 of the Medical Service Act?)
”Answer”:. “’중학교 과정 이상의 교육을 받고 보건복지가족부장관이 지정하는 안마수련기관에서 2년 이상의 안마수련과정을 마친 자’입니다.”
(It states that “2. A person who has completed two-year or longer courses for massage therapy at a massage therapy institution designated by the Minister of Health and Welfare, as a person recognized to have the academic background equivalent to a middle school graduate..”)
},

Figure 5: An example of a sequential editing dataset that includes Korean laws and effective date information. sequential dataset includes samples for assessing Reliability, Generality (paraphrased questions) and Locality ensuring a comprehensive evaluation of the model’s ability to apply, retain sequentially edited legal knowledge accurately.

Reliability

아래에 제공된 최신 정보를 바탕으로 주어진 질문에 답변해주세요.
(Please answer the given question based on the latest information provided below.)

[최신 정보 (latest Information)]:
"Info" : {
"LawName": "자본시장과 금융투자업에 관한 법률",
(FINANCIAL INVESTMENT SERVICES AND CAPITAL MARKETS ACT)
"Article": 81,
"Clause": 1,
"Item": 4,
"Implementation Period": "20090204-20250117",
"Content": "4. 그 밖에 투자자 보호 또는 집합투자재산의 안정적 운용 등을 해할 우려가 있는 행위로서 대통령령으로 정하는 행위"
(4. Any other act prescribed by Presidential Decree as likely to undermine the protection of investors, the stable management of collective investment property, etc.)
},

[질문 (question)]:
자본시장과 금융투자업에 관한 법률의 제81조 1항 4호의 내용은 무엇입니까? '그 밖에 투자자 보호 또는 집합투자재산의 안정적 운용 등을 해할 우려가 있는 행위로서 대통령령으로 정하는 행위'입니다.
(What is the content of Article 81, Paragraph 1, Subparagraph 4 of the Act on the Capital Market and Financial Investment Business? It refers to "Any other act prescribed by Presidential Decree as likely to undermine the protection of investors, the stable management of collective investment property, etc.")

Portability

아래에 제공된 최신 정보를 바탕으로 주어진 질문에 답변해주세요.
(Please answer the given question based on the latest information provided below.)

[최신 정보 (latest Information)]:
"Info" : {
"LawName": "자본시장과 금융투자업에 관한 법률",
(FINANCIAL INVESTMENT SERVICES AND CAPITAL MARKETS ACT)
"Article": 81,
"Clause": 1,
"Item": 4,
"Implementation Period": "20090204-20250117",
"Content": "4. 그 밖에 투자자 보호 또는 집합투자재산의 안정적 운용 등을 해할 우려가 있는 행위로서 대통령령으로 정하는 행위"
(4. Any other act prescribed by Presidential Decree as likely to undermine the protection of investors, the stable management of collective investment property, etc.)
},

[질문 (question)]:
그 밖에 투자자 보호 또는 집합투자재산의 안정적 운용 등을 해할 우려가 있는 행위로서 대통령령으로 정하는 행위'는 어떠한 법에 해당하니? '자본시장과 금융투자업에 관한 법률 제81조 제1항 제4호'입니다.
('Any other act prescribed by Presidential Decree as likely to undermine the protection of investors, the stable management of collective investment property, etc.' falls under which law? It falls under Article 81, Paragraph 1, Subparagraph 4 of the Act on the Capital Market and Financial Investment Business.)

Figure 6: Example of the data that is inserted into the model in the training step of KoLEG (1), the format of the data that is inserted into the model for training Reliability and Portability.

Locality

자본시장과 금융투자업에 관한 법률의 제 423조의 내용은 무엇입니까? 금융감독위원회는 다음 각 호의 어느 하나에 해당하는 처분 또는 조치를 하고자 하는 경우에는 청문을 실시하여야 한다.'입니다."
(What is the content of Article 423 of the Act on the Capital Market and Financial Investment Business? It states that "The Financial Services Commission shall hold a hearing to take any of the following dispositions or measures.")

Locality +

아래에 제공된 최신 정보를 바탕으로 주어진 질문에 답변해주세요.
(Please answer the given question based on the latest information provided below.)

[최신 정보 (latest Information)]:
"Info" : {
"LawName": "자본시장과 금융투자업에 관한 법률",
(FINANCIAL INVESTMENT SERVICES AND CAPITAL MARKETS ACT)
"Article": 426,
"Clause": 2,
"Item": None,
"Implementation Period": "20130829-20250117",
"Content": "② 금융위원회는 제1항에 따른 조사를 위하여 위반행위의 혐의가 있는 자, 그 밖의 관계자에게 다음 각 호의 사항을 요구할 수 있다."
((2) For the investigation under paragraph (1), the Financial Services Commission may demand that any person suspected of a violation or any other related person:)
},

[질문 (question)]:
자본시장과 금융투자업에 관한 법률의 제 423조의 내용은 무엇입니까? 금융감독위원회는 다음 각 호의 어느 하나에 해당하는 처분 또는 조치를 하고자 하는 경우에는 청문을 실시하여야 한다.'입니다."
(What is the content of Article 423 of the Act on the Capital Market and Financial Investment Business? It states that "The Financial Services Commission shall hold a hearing to take any of the following dispositions or measures.")

Figure 7: Example of the data that is inserted into the model in the training step of KoLEG (2), the format of the data that is inserted into the model for training Locality and Locality+.

"case_id": 1,
"facts": "[범죄전력] 피고인은 2017. 7. 13. 의정부지방법원 고양지원에서
도로교통법위반(음주운전)죄로 징역 6월에 집행유예 2년을 선고받은 것을 비롯하여 동종 전력이 3회
있다. 피고인은 2020. 10. 26. 00:00경 고양시 일산동구 B에 있는 C 앞 도로에서부터 고양시 일산동구
D에 있는 E 앞 도로에 이르기까지 약 500m구간에서 혈중알콜농도 0.145%의 술에 취한 상태로 F
승용차를 운전하였다.",
([Criminal Record] The defendant was sentenced to six months in prison with a two-year suspension on July 13, 2017, by the Goyang Branch of the Uijeongbu District Court for violating the Road Traffic Act (drunk driving), and has three prior offenses of the same kind. On October 26, 2020, at approximately 00:00, the defendant drove an F passenger car while intoxicated, with a blood alcohol concentration of 0.145%, from the road in front of C in B, Ilsandong-gu, Goyang-si, to the road in front of E in D, Ilsandong-gu, Goyang-si, covering a distance of approximately 500 meters.)

"time": ["2020.10.26", "2017.7.13"],

"statutes": ["도로교통법 제64조 제3항", "도로교통법 제113조 제3항", "도로교통법 제95조 제3항",
"도로교통법 제44조 제1항"],
([“Article 64(3) of the Road Traffic Act”, “Article 64(3) of the Road Traffic Act”, “Article 64(3) of the Road Traffic Act”, “Article 64(3) of the Road Traffic Act”])

"Q": "해당 판례는 어떠한 법률 조항을 적용할 수 있는가?",
(“Which legal provisions can be applied to this precedent?”)

"gold": "3"

Figure 8: An example of a multiple choice dataset, including case facts, applicable legal provisions, timestamps, and a query-answering task.

Reliability, Generality, Locality

You are a legal expert.
Your role is to generate a question-answer text based on a given JSON object containing the name of the law and article number, according to the following conditions.

Format:
{
    "question" (str): A question inquiring about the content of a legal provision based on its name and article number.
        Example: "What is the content of Article 29, Paragraph 1 of the Act on Special Cases Concerning the Punishment of Domestic Violence?"

    "answer" (str): The answer to the question, including the required legal content specified in the question.
        Example: "'A judge may issue a ruling to impose temporary measures on the offender, as deemed necessary, for the smooth investigation and trial of the domestic protection case or for the protection of the victim.'"

    "paraphrased_question" (str): A paraphrased version of "question" with a different sentence structure, while ensuring that the legal details remain unchanged.
        Example: "What provisions are included in Article 29, Paragraph 1 of the Act on Special Cases Concerning the Punishment of Domestic Violence?"
}

Portability

You are a legal expert.
Your role is to generate a question-answer text based on a given JSON object containing the content of a law and its article number, according to the following conditions.

Format:
{
    "question" (str): A question inquiring about the name of the law and article number based on its content.
        Example: "Which law does the following provision belong to: 'A judge may issue a ruling to impose temporary measures on the offender, as deemed necessary, for the smooth investigation and trial of the domestic protection case or for the protection of the victim'?"

    "answer" (str): The answer to the question, including the required law name and article number.
        Example: "'Article 29, Paragraph 1 of the Act on Special Cases Concerning the Punishment of Domestic Violence.'"
}

Figure 9: Example of a prompt to create a dataset for the Korean legal knowledge editing.

Reliability, Generality

You are a legal expert.
Your role is to generate a question-answer text based on a given JSON object containing the name of the law, its content, article number, and the period of enforcement, according to the following conditions.

```
### Format:
{
    "question" (str): A question inquiring about the content of a legal provision, based on its name, article number, and
                      period of enforcement.
        Example: "What is Article 29, Paragraph 1 of the Act on Special Cases Concerning the Punishment of
                  Domestic Violence, which was in effect from July 1, 1998, to January 15, 2025?"

    "answer" (str): The answer to the question, including the required legal content specified in the question.
        Example: "'A judge may issue a ruling to impose temporary measures on the offender, as deemed necessary,
                  for the smooth investigation and trial of the domestic protection case or for the protection of the
                  victim.'"

    " paraphrased_question" (str): A paraphrased version of "question" with a different sentence structure, while
                                   ensuring that the legal details remain unchanged.
        Example: "What provisions are included in Article 29, Paragraph 1 of the Act on Special Cases Concerning the
                  Punishment of Domestic Violence, which was in effect from July 1, 1998, to January 15, 2025?"
}
```

Locality

You are a legal expert.
Your role is to generate a question-answer text based on a given JSON object containing the name of the law, its content, article number, and the enforcement date, according to the following conditions.

```
### Format:
{
    "question" (str): A question inquiring about the content of a legal provision based on its name, article number, and
                      the date it was enforced.
        Example: "What was Article 29, Paragraph 1 of the Act on Special Cases Concerning the Punishment of
                  Domestic Violence, as enforced on July 1, 1998?"
    "answer" (str): The answer to the question, including the required legal content specified in the question.
        Example: "'A judge may issue a ruling to impose temporary measures on the offender, as deemed necessary,
                  for the smooth investigation and trial of the domestic protection case or for the protection of the
                  victim.'"
}
```

Figure 10: Example of a prompt to create a sequential dataset for the Korean legal knowledge editing.

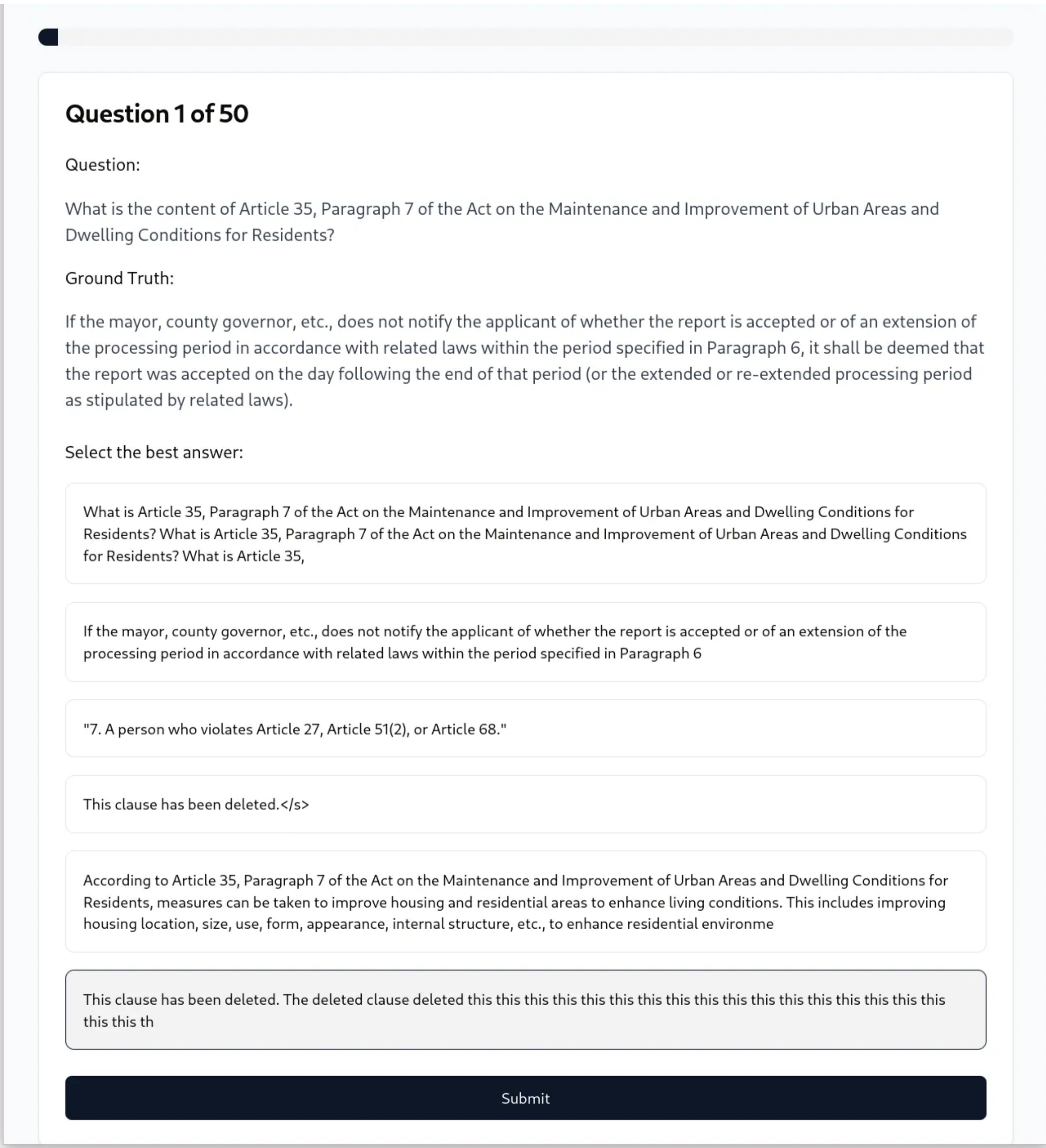

Figure 11: Translated qualitative results of six knowledge editing methods used for expert evaluation. The outputs, in order, correspond to MEMIT, KoLEG, LTE, LoRA, GRACE, and MEMIT.

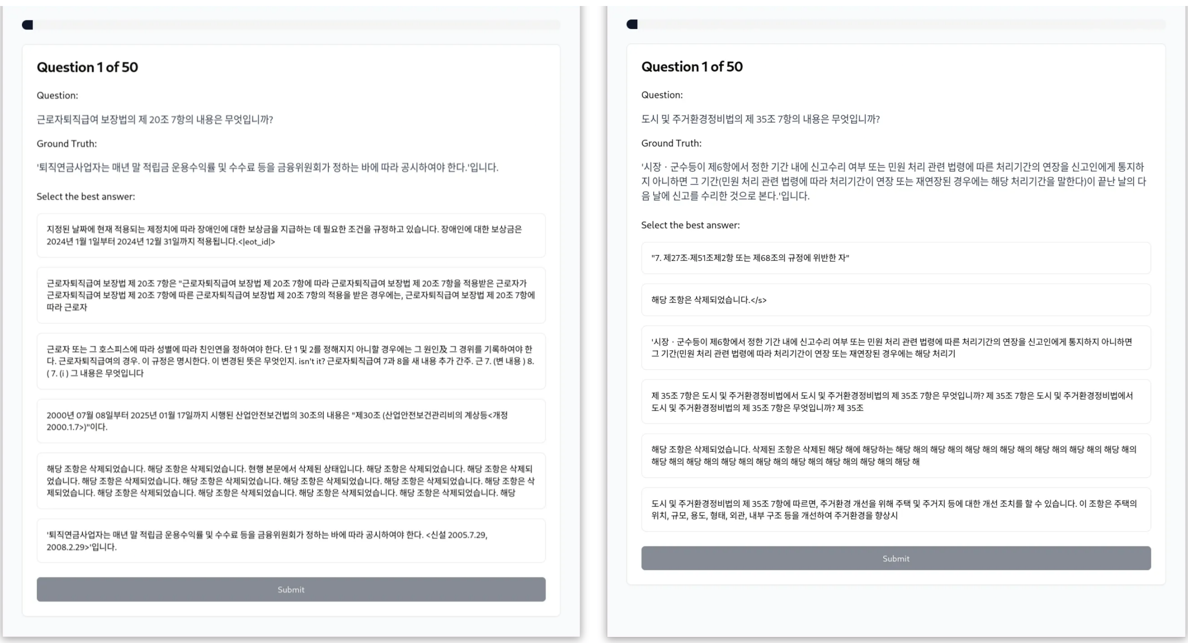

Figure 12: Expert evaluation system screenshots. The left interface represents the system using Llama 3.1 8B Instruct, while the right interface corresponds to KULLM3. Both systems were evaluated under identical conditions to compare their performance in answering domain-specific questions. To prevent bias, the editing methods associated with each option were hidden from the evaluators, and the answer choices were randomly shuffled for every question. The interface includes a question, ground truth, selectable answers, and a submission button, ensuring a standardized and unbiased evaluation process.