A Dual-Layer Knowledge Graph-Enhanced RAG Framework with Legal Equivalence Evaluation for Taiwanese Used Car Disputes

Bo-Zhi Yao
Department of Computer Science and Information Engineering
Fu Jen Catholic University
yao911001@gmail.com

Yi-Ting Peng*
Department of Computer Science and Information Engineering
Fu Jen Catholic University
164193@mail.fju.edu.tw

Abstract

The used car market suffers from significant information asymmetry, which frequently results in contractual disputes. Although large language models (LLMs) provide substantial potential for automated legal assistance, their application in legal domains remains challenging due to hallucination risks and difficulties in generating legally grounded responses. This study proposes a dual-layer Knowledge Graph Retrieval-Augmented Generation (KG-RAG) framework tailored specifically for Taiwanese used car disputes. We construct a structured legal knowledge graph integrating abstract statutory knowledge with historical judicial records collected between 2000 and 2020, thereby forming a litigation experience layer that complements the statutory knowledge layer. To prevent temporal data leakage, judicial records are chronologically partitioned, with cases from 2000 to 2020 used for knowledge graph construction and cases from 2021 to 2025 reserved for evaluation.

To address the semantic gap between colloquial user queries and formal legal representations, as well as the limitations of general-purpose semantic rerankers, our framework incorporates few-shot query enrichment, multi-hop graph expansion, and a weighted variant of the Reciprocal Rank Fusion (wRRF) strategy. Furthermore, we introduce a domain-specific "Soft Evaluation" framework based on a co-occurrence-based Legal Equivalence Mapping to complement strict exact-match evaluation. Comprehensive experiments show that the full KG-RAG configuration achieves a Hit@8 score of 62.81% under statute-level exact-match evaluation and 63.22% under soft-match evaluation. End-to-end generation evaluation further shows that the Qwen2.5-14B-based KG-RAG configuration achieves a generation hit rate of 55.79%, compared with 19.01% for the Qwen2.5-14B-only baseline and 43.80% for traditional vector-based RAG. The corresponding LTTR is 0.41%, compared with 11.16% for the Qwen2.5-14B-only baseline. These results show that the proposed framework improves legal retrieval and generation performance while substantially reducing high-risk recommendations relative to LLM-only generation.

Keywords: Knowledge Graph, Retrieval-Augmented Generation (RAG), Legal NLP, Large Language Models, Soft Evaluation, Loss-Trap Trigger Rate (LTTR)

1 Introduction

The used car market is characterized by severe information asymmetry, leading to contractual disputes involving concealed vehicle defects, mileage manipulation, and contract violations. Due to the complexity of legal procedures and statutory interpretation, ordinary consumers often have difficulty identifying appropriate legal claims and understanding potential remedies under legal frameworks such as Taiwan's Civil Code.

Recent advances in large language models (LLMs) have demonstrated strong capabilities in natural language understanding and knowledge-intensive tasks. However, directly applying general-purpose LLMs to legal applications remains challenging because of hallucination risks and unreliable legal reasoning. In particular, LLMs may generate plausible but incorrect legal explanations, including fabricated provisions or inappropriate interpretations of legal applicability, which significantly limits their deployment in high-stakes legal scenarios.

Retrieval-Augmented Generation (RAG) has emerged as an effective approach for improving the factual reliability of LLM-based systems by grounding generated responses in external knowl-

*Corresponding author.

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 86–94
November 6-7, 2026 ©2026 Association for Computational Linguistics

edge sources. Recent studies further demonstrate that integrating structured knowledge graphs into RAG pipelines can enhance multi-hop reasoning and improve retrieval effectiveness for complex knowledge-intensive tasks (Li et al., 2024; Peng et al., 2024; Saleh et al., 2024). However, legal retrieval remains particularly challenging because legal reasoning often depends not only on textual similarity but also on implicit relationships among factual circumstances, legal concepts, and judicial outcomes.

To address these challenges, this study proposes a dual-layer Knowledge Graph Retrieval-Augmented Generation (KG-RAG) architecture specifically designed for Taiwanese used car disputes. The proposed framework leverages historical judicial records collected between 2000 and 2020 to construct a litigation experience layer connected with abstract statutory knowledge. Unlike conventional semantic retrieval approaches, our framework employs multi-hop graph expansion to discover latent relationships between case facts and relevant legal provisions. Furthermore, a weighted variant of Reciprocal Rank Fusion (wRRF) is introduced to integrate retrieval signals from vector retrieval, knowledge graph traversal, and graph-based expansion.

In addition, existing retrieval evaluation methods based on strict exact matching may underestimate system performance in legal scenarios. Different statutory provisions may be closely related in practical litigation and may support overlapping legal claims. Motivated by this observation, we propose a domain-specific "Soft Evaluation" framework based on a co-occurrence-based Legal Equivalence Mapping as a complementary assessment alongside strict exact-match retrieval performance.

The contributions of this work are summarized as follows:

- We construct a dual-layer legal knowledge graph that integrates statutory knowledge with historical judicial experience for Taiwanese used car disputes.
- We propose a KG-RAG framework combining few-shot query enrichment, hybrid retrieval, multi-hop graph expansion, and a weighted variant of Reciprocal Rank Fusion (wRRF) for domain-specific legal reasoning.
- We introduce a co-occurrence-based soft evaluation mapping that complements strict exact-match metrics by accepting predefined related statutes for selected reference anchors.
- We incorporate litigation outcomes as retrieval weights and evaluate their contribution through progressive ablation studies.

2 Related Work

2.1 Structuring Unstructured Legal Texts

A primary challenge in legal information retrieval is the highly unstructured and scattered nature of raw litigation documents. Recent methodologies, such as StructRAG (Li et al., 2024), advocate for reconstructing documents into semi-structured knowledge representations to enhance retrieval signals for knowledge-intensive tasks. Following this paradigm, our framework does not rely solely on naive text chunking. Instead, it utilizes large language models (LLMs) to extract normalized entities—such as objective facts, legal concepts, and cited statutes—from raw judgments. These extracted elements are subsequently ingested into a Neo4j database to construct a structured dual-layer knowledge graph.

2.2 Query Enrichment for Short and Colloquial Inputs

In domain-specific applications, user queries are frequently short, colloquial, and lack explicit legal entities, creating a semantic mismatch between user expressions and formal legal representations. To address this issue, query enrichment techniques have been explored to bridge the semantic gap between user queries and domain-specific knowledge. Our framework adopts few-shot query enrichment to infer latent legal concepts from layperson queries. These extracted concepts are combined with the original input and used for semantic query expansion. This strategy improves the alignment between user intent and legal knowledge representations, thereby enhancing retrieval effectiveness.

2.3 Multi-hop Reasoning for Latent Legal Relationships

Complex legal inquiries often require reasoning across multiple nodes, as a single retrieval step may fail to capture all relevant legal relationships. Approaches such as KG-RAG with "Chain of Explorations" address this limitation by enabling step-by-step traversal along lengthy relational paths (Saleh

et al., 2024). Our framework adopts a similar multi-hop graph expansion strategy. By utilizing historical case nodes as bridging entities, the system enables retrieval of frequently associated and implicitly connected statutes (e.g., retrieving Article 359 alongside Article 354 of the Civil Code), which may remain undiscovered through standard vector retrieval alone.

2.4 Graph-Guided Retrieval Fusion for Legal Reasoning

Even when relevant candidates are retrieved, general-purpose embedding-based similarity ranking may fail to prioritize critical information in specialized domains. Recent studies highlight the importance of incorporating graph structures into ranking mechanisms (Dong et al., 2024), allowing retrieval systems to leverage document-level relationships during ranking. In our experiments, we observe that generalized neural rerankers may produce suboptimal ranking results in this specialized legal setting. Consequently, our framework adopts a weighted variant of the Reciprocal Rank Fusion (wRRF) strategy. Inspired by recent GraphRAG developments (Peng et al., 2024), this approach integrates signals from multiple retrieval sources and prioritizes legally relevant information supported by structural, semantic, and empirical litigation evidence.

3 Methodology

3.1 Dual-layer Knowledge Graph Design

To overcome the limitations of purely semantic retrieval, our system is built upon a dual-layer knowledge graph. Recent studies have shown that integrating structured knowledge representations with retrieval-augmented generation improves the reliability of knowledge-intensive reasoning tasks by providing explicit relational information beyond conventional text retrieval (Li et al., 2024; Sanmartín, 2024; Peng et al., 2024).

Our knowledge graph consists of two complementary layers:

- **Layer 1 (Legal Knowledge Layer):** This layer encodes abstract statutory knowledge, consisting of nodes representing articles of Taiwan's Civil Code and fundamental legal concepts. Similar legal KG construction approaches have demonstrated the effectiveness of representing legal rules, concepts, and relationships as structured graphs for improving legal reasoning and retrieval (Jain et al., 2022; Belikov and Raoult, 2025).
- **Layer 2 (Litigation Experience Layer):** This layer captures practical judicial experience, including historical case nodes, extracted factual elements, litigation outcomes, and cited legal provisions. Such experience-oriented representations allow retrieval systems to connect abstract legal norms with real-world judicial applications (Pham and Phan, 2026; Chen et al., 2026).

These two layers are connected through factual relationships, citation relationships, and legal claim associations, enabling the system to bridge abstract legal knowledge with practical litigation experience. Compared with traditional retrieval approaches that mainly rely on textual similarity, graph-based retrieval can exploit latent relationships among entities and improve multi-hop reasoning capability (Saleh et al., 2024; Dong et al., 2024).

To incorporate practical litigation experience into retrieval decisions, historical litigation outcomes are integrated into case nodes as outcome-aware weights. These weights are derived from judgment results, allowing different judicial outcomes to provide differentiated retrieval signals. Specifically, favorable, partially favorable, and unfavorable cases were assigned weights of 2.0, 1.2, and 0.8, respectively. Unfavorable results are assigned a positive weight (0.8) rather than zero or a negative value because, even when a plaintiff loses, the historical case still contains highly relevant factual descriptions and judicial reasoning pertinent to the legal dispute. The positive weight ensures these cases contribute to the retrieval context while being appropriately ranked below favorable precedents. These weights were empirically determined through preliminary experiments conducted exclusively on the knowledge construction set, without using any information from the evaluation set. The assigned weights modulate the contribution of historical cases during graph-based retrieval while preserving information from diverse judicial outcomes.

3.2 Infrastructure and Base Models

The proposed system is deployed locally to ensure data privacy and computational stability when processing sensitive legal documents. Privacy-preserving processing is particularly important in

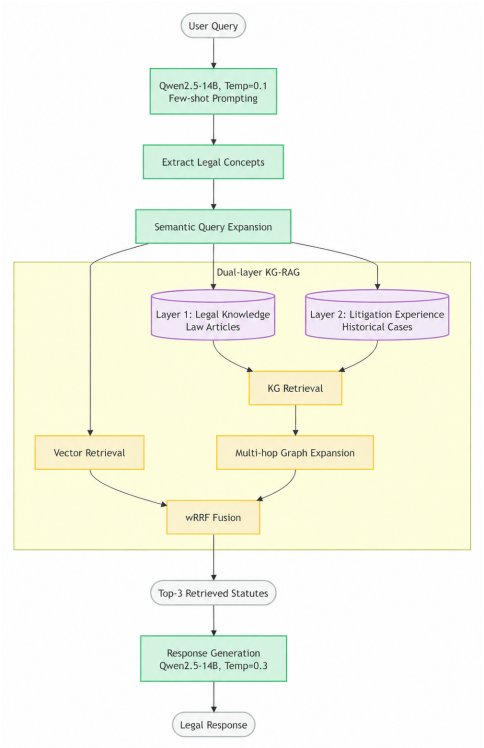

Figure 1: Overview of the proposed dual-layer KG-RAG framework.

legal applications because judicial documents frequently contain sensitive personal information. Existing research has emphasized the necessity of reliable de-identification mechanisms when constructing legal datasets (Larson et al., 2024).

The proposed system employs `Qwen2.5-14B` through the Ollama framework for query generation, statute extraction, relevance classification, and its primary end-to-end response generation configuration. The underlying knowledge graph is managed by `Neo4j`. Cypher queries are employed to navigate relationships among case rulings, factual elements, and legal statutes. For text embedding, the `BGE-M3` model is adopted to map colloquial user queries and legal concepts into a shared high-dimensional embedding space. For the end-to-end comparative evaluation, `Llama3` is additionally included as an alternative base LLM under the same no-RAG, traditional RAG, and KG-RAG settings.

Recent legal QA studies have shown that combining large language models with retrieval mechanisms can improve factual grounding and reduce hallucination risks in high-stakes domains (Reuter et al., 2025; Li et al., 2026).

3.3 The KG-RAG Pipeline

The retrieval pipeline consists of three critical stages:

1. **Query Formulation and Semantic Expansion:**

   To bridge the semantic gap between layperson language and formal legal terminology, the system first extracts latent legal concepts and intents using few-shot prompting. Query enrichment has been widely explored as an effective strategy for improving retrieval performance when user queries are short, ambiguous, or lack domain-specific terminology (Perna, 2025).

   A prompt-based constraint is incorporated into the query enrichment stage to prevent the LLM from predicting specific statute numbers, thereby reducing early-stage hallucination risks. These extracted concepts are then combined with the original query and used for semantic query expansion.

2. **Tri-Track Hybrid Retrieval:**

   The retrieval process operates across three parallel tracks:

   (1) Vector-based Semantic Retrieval, which

identifies contextually similar nodes based on embedding similarity;

(2) Knowledge Graph-based Retrieval, which matches extracted entities and traverses legal and factual relationships within the knowledge graph;

(3) Multi-hop Graph Expansion, which utilizes retrieved seed nodes to identify highly co-occurring statutes through multi-hop graph traversal.

Multi-hop graph reasoning has been shown to improve retrieval effectiveness in complex question answering tasks by enabling systems to discover indirect relationships that cannot be captured through single-step retrieval (Saleh et al., 2024; Lee et al., 2025).

3. **wRRF:**

Our preliminary experiments indicated that direct embedding-based similarity ranking using BGE-M3 resulted in domain-specific ranking degradation. This degradation may result from the discrepancy between general semantic similarity and legal relevance, where semantically similar provisions may not necessarily represent equivalent legal strategies. Therefore, we adopt a weighted variant of Reciprocal Rank Fusion (wRRF) to aggregate ranking signals from the tri-track retrieval framework.

Graph-guided ranking methods have demonstrated that incorporating structural information can improve retrieval ranking quality compared with purely semantic approaches (Dong et al., 2024). Inspired by recent GraphRAG developments (Peng et al., 2024), this approach integrates signals from multiple retrieval sources and prioritizes legally relevant information supported by structural, semantic, and empirical litigation evidence.

$$\text{Score}(d) = w_v\, R_v(d) + w_g\, R_g(d) + w_e\, R_e(d)$$

where the coefficients were empirically set to $w_v = 1.0$, $w_g = 1.5$, and $w_e = 1.2$ based on preliminary experiments conducted on the knowledge construction set. These coefficients were fixed before evaluation.

Each reciprocal rank score is computed as

$$R_i(d) = \frac{1}{k + r_i(d)}, \qquad i \in \{v, g, e\},$$

where $r_i(d)$ denotes the ranking position of document d in retrieval track i, and k is the smoothing constant used in Reciprocal Rank Fusion. In all experiments, we fixed $k = 60$, following the standard Reciprocal Rank Fusion formulation.

3.4 Soft Evaluation via Legal Equivalence

Traditional retrieval evaluation metrics usually rely on exact matching between retrieved documents and reference answers. However, such evaluation strategies may underestimate system capability in legal reasoning scenarios because multiple legal provisions can be closely related in the same litigation strategy or claim structure.

Recent RAG evaluation studies have highlighted the limitations of lexical matching metrics and emphasized the importance of semantic and task-oriented evaluation frameworks (Yu et al., 2024; Balamurali and Cheng, 2025). Similarly, legal benchmarks have demonstrated that legal reasoning requires evaluation beyond simple textual overlap (Joshi et al., 2024).

Therefore, we define a co-occurrence-based "Legal Equivalence Mapping" from statute relationships observed in the historical judgments used for knowledge graph construction. Neo4j is used to identify provisions that frequently co-occur in the same historical cases. The resulting relations are encoded as a fixed, reference-anchored lookup table before evaluation; no information from the 2021–2025 evaluation set is used to construct the mapping. The mapping contains five anchor statutes and twelve unique statutes in total. For a given reference statute, soft matching expands the acceptable target set only when that reference statute appears as an anchor in the mapping. The implementation is therefore directional and should not be interpreted as a set of disjoint or transitively closed legal clusters.

Anchor statute	Additional statutes accepted under soft matching
Civil Code Art. 354	Arts. 359, 360, 356, 365, 259
Civil Code Art. 359	Arts. 354, 259, 227
Civil Code Art. 349	Art. 353
Civil Code Art. 345	Art. 367
Civil Code Art. 254	Art. 259

Table 1: Reference-anchored Legal Equivalence Mapping used only for soft-match evaluation.

Soft Evaluation does not modify retrieval outputs or ranking scores; it changes only the scoring criterion after retrieval. Under exact matching, only an identical reference statute is counted as a hit. Under soft matching, the retrieved list is also compared with the additional statutes associated with each reference anchor in Table 1. Because this mapping was derived from corpus-level co-occurrence analysis rather than independently annotated by multiple human annotators, inter-annotator agreement is not reported. This is a limitation of the current evaluation design, and future work will investigate expert-validated equivalence relations with multiple annotators.

This complementary evaluation provides an additional view of retrieval performance while preserving exact-match Hit@8 as the primary strict metric.

4 Experiments and Results

4.1 Experimental Setup

This study utilizes a custom-built dataset comprising publicly available Taiwanese civil judgments related to used car disputes, collected from the Judicial Yuan's public judgment database.

The dataset is temporally partitioned to prevent data leakage and simulate real-world prospective retrieval. The knowledge construction set consists of 444 historical judgments collected from 2000 to 2020 and was used exclusively for knowledge graph construction, while the final evaluation set contains 242 recent judgments collected from 2021 to 2025 and was reserved exclusively for performance assessment.

No judgments from the evaluation period were used during knowledge graph construction or parameter selection to prevent temporal data leakage. This chronological partition ensures that all retrieval models and hyperparameters are determined without access to future judgments.

To filter out irrelevant records, we employed a locally hosted Qwen2.5-14B model with deterministic decoding (temperature = 0.0) to perform LLM-based semantic relevance classification. The model was instructed to return a structured JSON object containing a binary relevance label together with a brief justification. Cases classified as irrelevant, along with the generated justifications, were retained for subsequent manual inspection. Strict rule-based de-identification algorithms were applied to mask Personally Identifiable Information (PII). Similar privacy-aware preprocessing and de-identification strategies have been explored in previous legal dataset construction studies (Larson et al., 2024).

Finally, the extracted factual descriptions were transformed into colloquial user queries representing realistic legal inquiries. Each judgment was converted into one corresponding user query using `Qwen2.5-14B` with a fixed prompting template. The same model was also used to extract candidate statutes from each full judgment. To reduce unsupported statute references, each extracted statute was then normalized and automatically verified against the complete judgment text; statutes whose textual forms did not occur in the judgment were removed. A random manual audit of 20 evaluation cases found that all 20 cases (100%) had correctly extracted and text-verified reference statutes. Nevertheless, because both query generation and statute extraction are LLM-assisted, the benchmark should be interpreted as an LLM-assisted evaluation set rather than a fully human-annotated gold standard. Consequently, each judgment is associated with exactly one generated colloquial query and one or more verified reference statutes.

All experiments were executed on a local GPU environment equipped with an NVIDIA RTX 5080.

4.2 Knowledge Graph Construction

The foundational legal knowledge graph was constructed by parsing Markdown-formatted statutory documents, including Taiwan's Civil Code provisions and related legal concepts.

Subsequently, the 444 historical judgments from the knowledge construction set were processed to extract objective facts, legal issues, litigation outcomes, and cited statutes. Knowledge graph construction from legal documents has been explored in previous studies as an effective method for transforming unstructured judicial texts into structured reasoning resources (Belikov and Raoult, 2025; Jain et al., 2022).

To incorporate practical litigation experience into retrieval, historical case nodes were augmented with outcome-aware weights according to judicial results. Specifically, favorable, partially favorable, and unfavorable cases were assigned weights of 2.0, 1.2, and 0.8, respectively.

These elements were mapped into the Neo4j database as interconnected nodes and directed

edges, including HAS_FACT, HAS_ISSUE, and CITES relationships.

4.3 Retrieval Evaluation and Ablation Studies

To evaluate retrieval performance, we conducted progressive ablation studies on the 242-case evaluation set. Each query is associated with one or more verified reference statutes, which serve as the retrieval targets. Retrieval performance is measured using Hit@8. Under exact matching, a query is counted as a hit when at least one reference statute is retrieved within the top-8 ranked candidates. Under soft matching, the acceptable target set is expanded according to the fixed reference-anchored Legal Equivalence Mapping in Table 1; a query is counted as a soft hit when at least one retrieved statute matches either an exact reference statute or one of the additional statutes associated with that reference anchor.

Model Configuration	Hit@8 (Exact) (%)	Hit@8 (Soft) (%)
A. Vector-only	50.41	54.13
B. Vector + Few-shot (Query Enrichment)	50.41	53.31
C. Vector + Graph (KG Retrieval)	56.61	58.26
D. Vector + Graph + Multi-hop	58.26	58.68
E. Vector + Graph + Multi-hop + Outcome Weight	58.26	58.68
F. Full KG-RAG + wRRF	**62.81**	**63.22**

Table 2: Hit@8 retrieval performance under exact-match and soft-match evaluation.

As shown in Table 2, the vector-only baseline achieved an exact Hit@8 of 50.41% and a soft Hit@8 of 54.13%. Adding few-shot query enrichment alone did not change exact Hit@8 and slightly decreased soft Hit@8 to 53.31%. In contrast, adding graph retrieval increased exact Hit@8 to 56.61% and soft Hit@8 to 58.26%, while adding multi-hop expansion further increased exact Hit@8 to 58.26% and soft Hit@8 to 58.68%. Adding outcome weights to the multi-hop configuration produced the same Hit@8 values under both metrics, indicating no additional gain under this retrieval metric in isolation. The full KG-RAG configuration with wRRF achieved the highest observed retrieval scores in this evaluation, reaching 62.81% exact Hit@8 and 63.22% soft Hit@8.

4.4 End-to-End Generation and Risk Mitigation

Although retrieval performance is important, practical legal assistance requires not only retrieving relevant provisions but also generating reliable and appropriately grounded responses.

Therefore, we further evaluate end-to-end generation performance using two metrics: Generation Hit Rate and the Loss-Trap Trigger Rate (LTTR).

For each evaluation query, the verified statutes associated with the corresponding original judgment served as the reference set. A generated response was counted as a generation hit if it recommended at least one statute consistent with the reference set. This metric therefore measures whether the response identifies at least one supported statutory provision rather than whether the full legal explanation is completely correct.

Evaluation of RAG systems has increasingly emphasized response reliability and factual grounding rather than retrieval accuracy alone (Yu et al., 2024; Reuter et al., 2025).

LTTR measures the frequency with which generated responses recommend high-risk legal claims that are frequently raised by laypersons but demonstrate relatively low success rates according to historical litigation outcomes.

Formally, the Loss-Trap Trigger Rate (LTTR) is defined as:

$$\text{LTTR} = \frac{N_{\text{high-risk}}}{N_{\text{queries}}} \times 100\%$$

where $N_{\text{high-risk}}$ denotes the number of generated responses that trigger at least one high-risk legal claim, and N_{queries} represents the total number of evaluated queries.

In this study, legal provisions with historical success rates below an empirically defined threshold of 30% are categorized as high-risk claims. The threshold was selected based on the empirical distribution of claim success rates in the knowledge construction set and was fixed before evaluation. This threshold was chosen to identify legal claims with substantially lower historical success probability rather than to represent a universal legal risk boundary.

System Configuration	Generation Hit (%)	LTTR (%)
Llama3	5.79	1.65
Qwen2.5-14B	19.01	11.16
Llama3 + Traditional RAG (Vector)	36.78	0.00
Qwen2.5-14B + Traditional RAG (Vector)	43.80	0.00
Llama3 + Dual-layer KG-RAG	45.04	0.41
Qwen2.5-14B + Dual-layer KG-RAG	**55.79**	**0.41**

Table 3: End-to-end generation performance on the 242-case evaluation set across two base LLMs.

As shown in Table 3, the same overall generation trend is observed across both base LLMs. For Llama3, the generation hit rate increased from

5.79% without retrieval to 36.78% with traditional vector-based RAG and 45.04% with the dual-layer KG-RAG configuration. For Qwen2.5-14B, the corresponding generation hit rates were 19.01%, 43.80%, and 55.79%, respectively. Traditional vector-based RAG achieved an LTTR of 0.00% for both base models, while the dual-layer KG-RAG configuration achieved 0.41% for both. The LLM-only baselines produced LTTR values of 1.65% for Llama3 and 11.16% for Qwen2.5-14B. These results show that retrieval substantially improves statute generation hit rates for both base LLMs, while traditional vector-based RAG yields the lowest LTTR in this evaluation.

The retrieval and end-to-end results should be interpreted together. Retrieval performance is measured against verified statute references, whereas generation hit rate only requires the final response to identify at least one supported statute. The two metrics therefore capture different stages of the pipeline and are not expected to have identical numerical values.

5 Conclusion and Future Work

5.1 Conclusion

This study proposes a dual-layer KG-RAG system specifically designed for used car disputes in Taiwan. By integrating multi-hop graph retrieval with a weighted variant of Reciprocal Rank Fusion (wRRF), the proposed framework combines vector, graph, and expansion-based retrieval signals. The retrieval ablation study shows that the full KG-RAG configuration achieves 62.81% exact Hit@8 and 63.22% soft Hit@8 on the 242-case evaluation set.

The end-to-end evaluation with Qwen2.5-14B shows a generation hit rate of 55.79% for the dual-layer KG-RAG configuration, compared with 43.80% for traditional vector-based RAG and 19.01% for the LLM-only baseline. The corresponding LTTR values are 0.41%, 0.00%, and 11.16%, respectively. A consistent generation trend is also observed with Llama3, for which the generation hit rate increases from 5.79% to 36.78% with traditional RAG and 45.04% with KG-RAG. These results indicate that graph-based retrieval improves the probability of producing a response that cites at least one supported statute across both evaluated base LLMs. At the same time, the lower LTTR of traditional RAG shows that retrieval accuracy and risk control are distinct evaluation dimensions that should be reported together. Rather than replacing legal professionals, the proposed framework is designed as an evidence-based legal assistance system that assists non-expert users in understanding potential legal options before seeking professional consultation.

5.2 Future Work

Future work will focus on several research directions. First, we aim to extend the underlying knowledge graph beyond used car disputes to cover a broader range of consumer-related legal scenarios, enabling the proposed framework to support everyday legal inquiries. Second, to maintain the relevance and accuracy of the knowledge graph, an automated pipeline for continuously incorporating newly published court decisions and statutory amendments will be investigated.

Third, future work will strengthen the Soft Evaluation protocol by validating the current co-occurrence-based mapping with multiple legal annotators and reporting inter-annotator agreement. We will also investigate whether symmetric or graph-derived equivalence relations provide a more principled alternative to the current directional reference-anchored mapping. Finally, comprehensive user studies with both non-expert users and legal professionals will evaluate the usability, interpretability, and practical effectiveness of KG-RAG-based legal assistance systems.

Acknowledgment

This work was supported by the National Science and Technology Council, Taiwan, under Grant NSTC 115-2221-E-030-016.

References

Sai Shridhar Balamurali and Lu Cheng. 2025. Revisiting nli: Towards cost-effective and human-aligned metrics for evaluating llms in question answering. *arXiv preprint arXiv:2511.07659*.

Alexander V. Belikov and Sacha Raoult. 2025. Knowledge graphs construction from criminal court appeals: Insights from the french cassation court. *arXiv preprint arXiv:2501.14579*.

Zerui Chen, Qinggang Zhang, Zhishang Xiang, Zhimin Wei, Linfeng Gao, Xiao Huang, Zhihong Zhang, and Jinsong Su. 2026. Legalgraphrag: Multi-agent graph retrieval-augmented generation for reliable legal reasoning. In *Proceedings of the 64th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 37455–37484. Association for Computational Linguistics.

Jialin Dong, Bahare Fatemi, Bryan Perozzi, Lin F. Yang, and Anton Tsitsulin. 2024. Don't forget to connect! improving rag with graph-based reranking. *arXiv preprint arXiv:2405.18414*.

Sarika Jain, Pooja Harde, Nandana Mihindukulasooriya, Sudipto Ghosh, Abhinav Dubey, and Ankush Bisht. 2022. Constructing a knowledge graph from indian legal domain corpus. In *Proceedings of the TEXT2KG Workshop at ESWC*.

Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. Il-tur: Benchmark for indian legal text understanding and reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11460–11499. Association for Computational Linguistics.

Stefan Larson, Nicole Cornehl Lima, Santiago Pedroza Diaz, Amogh Manoj Joshi, Siddharth Betala, Jamiu Tunde Suleiman, Yash Mathur, Kaushal Kumar Prajapati, Ramla Alakraa, Junjie Shen, Temi Okotore, and Kevin Leach. 2024. De-identification of sensitive personal data in datasets derived from iit-cdip. In *Proceedings of EMNLP*.

Jihyung Lee, Daehui Kim, Seonjeong Hwang, Hyounghun Kim, and Gary Geunbae Lee. 2025. Open legal question answering with multi-hop reasoning. In *Proceedings of EMNLP*.

Yaocong Li, Qiang Lan, Leihan Zhang, and Le Zhang. 2026. Legal-dc: Benchmarking retrieval-augmented generation for legal documents. *arXiv preprint arXiv:2603.11772*.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2024. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. *arXiv preprint arXiv:2410.08815*.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*.

Massimo Perna. 2025. Knowledge graph for query enrichment in retrieval augmented generation in domain specific application. Master's thesis, University of Twente.

Vuong T. Pham and Minh Phan. 2026. Integrating knowledge graph with retrieval-augmented generation for vietnamese legal question answering. *DS Journal of Digital Science and Technology*, 5(1):1–13.

Markus Reuter, Tobias Lingenberg, Ruta Liepina, Francesca Lagioia, Marco Lippi, Giovanni Sartor, Andrea Passerini, and Burcu Sayin. 2025. Towards reliable retrieval in rag systems for large legal datasets. In *Proceedings of NLLP*.

Ahmmad O. M. Saleh, Gokhan Tur, and Yucel Saygin. 2024. Sg-rag: Multi-hop question answering with large language models through knowledge graphs. In *Proceedings of ICNLSP*.

Diego Sanmartín. 2024. Kg-rag: Bridging the gap between knowledge and creativity. *arXiv preprint arXiv:2405.12035*.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. *arXiv preprint arXiv:2405.07437*.