Retrieving the Right Law: Enhancing Legal Search with Style Translation

Szu-Ju Chen*
b10705005@csie.ntu.edu.tw
Department of Computer Science and Information Engineering, National Taiwan University
Taipei, Taiwan

Jing Jin*
b10204022@ntu.edu.tw
Department of Computer Science and Information Engineering, National Taiwan University
Taipei, Taiwan

Sheng-Lun Wei
weisl@nlg.csie.ntu.edu.tw
Department of Computer Science and Information Engineering, National Taiwan University
Taipei, Taiwan

Chien-Hung Chen
chchen@nlg.csie.ntu.edu.tw
Graduate Institute of Networking and Multimedia, National Taiwan University
Taipei, Taiwan

Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University
Taipei, Taiwan

Abstract

Legal question answering requires accurate retrieval of relevant laws, yet the significant writing style gap between user queries and legal provisions poses a major challenge. Existing datasets and retrieval methods often struggle to capture the complexity of legal language, limiting retrieval effectiveness. In this study, we introduce the Legal Query-to-Provision Retrieval (LQPR) task and construct Query2Provision (Q2P), a dataset designed to enhance law retrieval by incorporating diverse case scenarios and linguistic structures representative of real-world legal inquiries. To address the style disparity, we propose a style translation approach that transforms informal user queries into a more formal legal tone and simplifies complex legal provisions for better alignment. Our experiments demonstrate that integrating writing style transformation significantly improves retrieval performance. The dataset is available at https://github.com/ntunlplab/Query2Provision

CCS Concepts

• **Information systems** → **Document filtering**; **Question answering**.

Keywords

Legal Retrieval, Writing Style Adaptation, Legal Retrieval Dataset

ACM Reference Format:
Szu-Ju Chen, Jing Jin, Sheng-Lun Wei, Chien-Hung Chen, and Hsin-Hsi Chen. 2025. Retrieving the Right Law: Enhancing Legal Search with Style Translation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730246

*Both authors contributed equally to this research.

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730246

1 INTRODUCTION

Legal issues frequently arise in everyday life, yet legal texts are often complex and require domain-specific knowledge to interpret. This makes it difficult for the general public to understand and apply relevant laws to their situations [10]. Moreover, when faced with a legal question, individuals often struggle to identify which specific laws or legal provisions are applicable to their case [1].

Knowing the correct legal references is crucial, as it allows individuals to interpret the law themselves or seek assistance from legal models for further explanation. For large language models (LLMs), having access to precise legal references is essential for generating accurate and reliable responses [15]. Additionally, providing transparency regarding which laws the model relies on enhances its interpretability, fostering greater trust in legal AI systems [10].

However, retrieving the correct legal provisions is challenging. It not only requires legal expertise but also faces the difficulty that legal texts and public queries are often expressed in significantly different styles [12]. For instance, while a person might ask, "*Can my landlord evict me without notice?*", [1] while the corresponding legal provision may be phrased in highly technical terms, such as: "*A landlord may not unilaterally terminate a lease agreement and require the tenant to vacate without proper notice and compliance with the procedures set forth in this Act.*" Compared to the concise and direct phrasing of a public inquiry, legal provisions are often structured formally, using terminology such as "proper notice" and "compliance with procedures." These stylistic differences significantly increase the difficulty of retrieving relevant provisions, as conventional search methods struggle to bridge this linguistic gap.

Additionally, public legal inquiries vary widely in form [1]. Some are short and direct, such as "*Is hit-and-run a prosecutable offense?*", while others resemble long-form case descriptions, where individuals narrate complex situations before posing a legal question. For

[1]The examples and prompts in this study are written in Chinese originally. For readability, we all translate them into English.

example: "*My husband and I rented a subdivided apartment. ...(500 words)... Should we file a lawsuit, or should we just take our two months' deposit and look for another place?*" Such variations further complicate legal retrieval, as models must not only handle diverse writing styles but also extract relevant legal concerns from lengthy, unstructured narratives.

We introduce the Legal Query-to-Provision Retrieval (LQPR) task, which focuses on retrieving legal provisions despite the significant stylistic differences between user queries and legal texts, making the task more challenging. To address this challenge, we construct a Query2Provision (Q2P) dataset of 1,375 legal query-to-provisions pairs collected from online legal QA platforms. This dataset maps informal, colloquial legal inquiries to their corresponding legal provisions across 299 legal codes, providing a valuable resource for studying legal retrieval in real-world scenarios. In our training data, each question corresponds to an average of three legal provisions, with a maximum of 18 provisions, indicating that legal questions often involve multiple regulations, increasing the complexity of retrieval.

Furthermore, we propose a style translation approach that effectively improves the performance of retrieval models in identifying relevant legal provisions. By bridging the linguistic gap between public legal questions and formal legal texts, our method enhances law retrieval performance, offering a significant step forward in legal retrieval research.

This study makes the following key contributions: a) We introduce the Legal Query-to-Provision Retrieval (LQPR) task, which requires not only legal knowledge but also techniques to overcome significant writing style differences between public queries and legal texts. b) We construct a Query2Provision (Q2P) dataset, sourced from legal QA platforms, which includes diverse question formulations and covers a broad range of legal provisions. c) We propose a style translation method and demonstrate through experiments and discussions that writing style transformation effectively enhances retrieval model performance.

2 RELATED WORK

Legal retrieval has been widely studied. Recently, much of the prior research focusing on legal case retrieval (LCR). The goal of LCR is to retrieve similar historical cases from a database given a query case [6]. This task is crucial for legal professionals, as past rulings serve as key references for legal reasoning and decision-making.

LCR shares several key challenges with legal statute retrieval. Both tasks require domain expertise, involve searching across multiple legal documents, must account for evolving legal frameworks, and demand highly trustworthy retrieval systems to ensure accurate and reliable legal references [7, 11]. However, a fundamental difference exists between LCR and statute retrieval: writing style similarity. In LCR, both queries and documents originate from legal professionals, resulting in a relatively consistent writing style. In contrast, legal statute retrieval faces a significant writing style gap, as user queries are often informal and unstructured, while legal provisions are written in highly technical and formal language. This discrepancy makes traditional retrieval methods less effective, necessitating specialized techniques to bridge the linguistic divide.

Table 1: Statistics of provisions and codes in Q2P dataset

Dataset	Provisions		Unique Codes	
	Mean	Max	Mean	Max
Training Set	3.25	18	1.61	7
Test Set	2.99	11	1.61	7

Existing datasets in this area include STARD[3], SLARD[13], and BSARD[9]. STARD and SLARD are based on Chinese legal systems, with STARD emphasizing real-world queries from the general public and SLARD targeting the retrieval of superior legal provisions across jurisdictions. BSARD, in contrast, is centered on the French legal system and includes expert-refined queries paired with articles from Belgian law. Compared to these datasets, our dataset is built from authentic interactions between laypersons and legal professionals within a Mandarin-speaking legal system.

3 DATASET CONSTRUCTION

3.1 LQPR Task Definition

Given a user query q and a set of legal provisions $S = \{s_1, s_2, ..., s_n\}$, the goal of the legal question-to-statute retrieval task is to identify the most relevant provisions $s^* \in S$ that correspond to the query q.

3.2 Data Collection

Our dataset is derived from Legispedia, [2] a legal knowledge-sharing platform dedicated to making legal information accessible to the general public. On this platform, users can post legal inquiries, which are answered by verified legal professionals. As part of their responses, these legal experts frequently reference specific legal provisions relevant to the questions being asked.

For the purpose of this study, we consider the referenced provisions included in the expert responses as the relevant legal provisions associated with each user query. We collected a total of 1,375 question-answer pairs, where each response explicitly cites one or more legal provisions. The referenced provisions were manually extracted and normalized to ensure consistency.

3.3 Statistics

Given the hierarchical structure of legal documents, we consider individual articles (e.g., Article 2 of the Civil Code) as the atomic units of retrieval, as we believe this granularity better preserves essential contextual information. For experimentation, we split the dataset into 1,175 training instances and 200 test instances to facilitate model evaluation. Our dataset covers 299 distinct legal codes, making it a diverse and comprehensive resource. Since the referenced legal provisions for a single query often originate from multiple legal documents, this task inherently aligns with cross-document information retrieval.

Table 1 presents the statistics of the extracted legal provisions and unique legal codes in our dataset. On average, each question references more than two legal provisions and spans more than one distinct legal code, highlighting the complexity of LQPR task. Additionally, our dataset has an average length of 152 words, with

[2] https://www.legis-pedia.com/

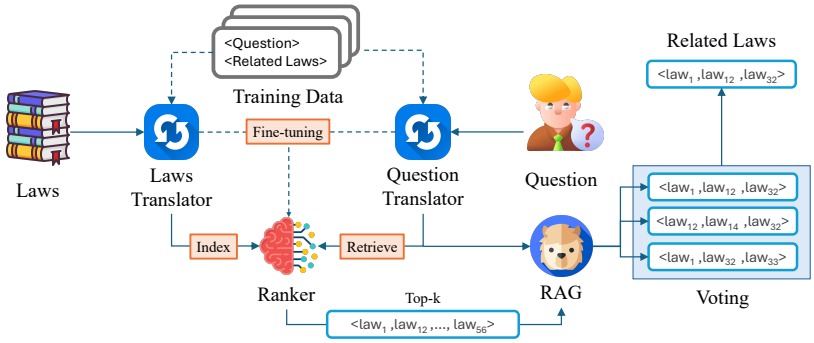

Figure 1: Our Legal Query-to-Provision Retrieval System with Style Translation

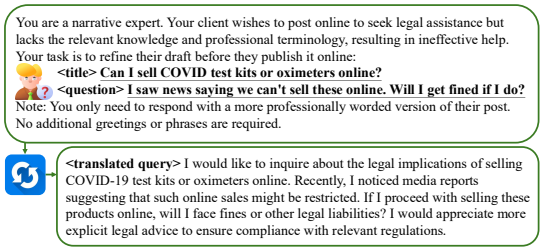

Figure 2: Prompt and Example of Query Translation

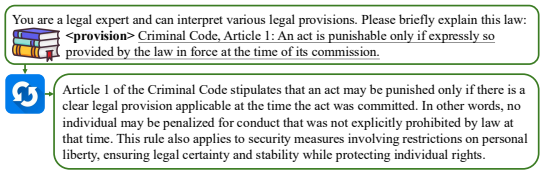

Figure 3: Prompt and Example of Provision Translation

a maximum of 623 words, highlighting the variability in legal query complexity.

4 METHODOLOGY

4.1 Style Translation Models

The Style Translation Models are responsible for transforming text into different writing styles, ensuring better alignment between user queries and legal provisions. This transformation is achieved by providing a style-specific prompt to a LLM, which then generates the reformulated text. Different prompts are used for queries and legal provisions to account for their distinct linguistic characteristics. The query transformation prompt shown in Figure 2 is designed to rephrase informal user questions into a formal legal tone, making them more compatible with the language used in legal provisions. The prompt for provision transformation (illustrated in Figure 3) simplifies complex legal language into an easier-to-understand form, making provisions more accessible to non-experts. By leveraging these style translation models, we bridge the stylistic gap between legal texts and user queries, enhancing the effectiveness of legal retrieval.

4.2 Ranker Fine-tuning

Inspired by BSARD [9], we adopt a BiEncoder as our ranker. This approach maps both queries and legal provisions into dense vector representations and computes their similarity scores to retrieve the most relevant provisions. After applying style translation, the transformed queries and provisions are used to fine-tune the ranker, allowing it to better capture the relationship between reformulated user questions and legal texts. The trained ranker is then employed for both indexing and retrieval, enabling efficient and accurate matching between user queries and legal provisions.

4.3 Majority Voting with RAG

After the Ranker retrieves the top-k most relevant legal provisions for a given query, these provisions are then passed to a LLM for retrieval-augmented generation (RAG). The LLM generates n possible answers, each suggesting relevant legal provisions based on the retrieved context. To improve reliability, we apply a majority voting strategy, selecting provisions that appear in at least m of the n generated results. These frequently occurring provisions are considered the final set of potentially relevant provisions, enhancing the robustness of the retrieval process.

5 EXPERIMENTS

5.1 Experiment Setup

In our experiment, we employ `GPT-4o` as the style translator to refine the textual input. For ranking, we utilize QLoRA [5] to fine-tune Chinese RoBERTa [3] [4] as the Ranker. The ranking process selects the top 15 candidates (top-k = 15) to ensure high recall while maintaining relevance.

5.2 Evaluation Metrics

Since retrieving too many irrelevant provisions or too few relevant ones can negatively impact downstream task performance and reduce interpretability, we evaluate our approach using Precision, Recall, and F1 Score in this study. Precision is calculated as the ratio of the number of correct predictions to the total number of predictions made by the model on the test dataset. It measures the ability of the model in retrieving only relevant legal provisions. Recall is calculated by dividing the number of correct predictions by

[3] hfl/chinese-roberta-wwm-ext

Table 2: Experimental Results

Models	Original			Style Translated		
	Precision	Recall	F1-score	Precision	Recall	F1-score
BM25	0.0605	0.0602	0.0604	0.1169	0.1171	**0.1170**
Ranker	0.1743	0.1758	0.1750	0.1827	0.1843	**0.1835**
LLaMA	0.1378	0.1438	0.1408	0.1552	0.2007	**0.1751**
LLaMA_voting	0.1692	0.1579	0.1633	0.1807	0.1967	**0.1884**
TaiwanLLaMA	0.2158	0.1555	0.1808	0.2470	0.1722	**0.2030**
TaiwanLLaMA_voting	0.2185	0.1943	0.2057	0.2284	0.2099	**0.2188**

the total number of relevant provisions labeled in the test dataset. This metric assesses the model's ability to identify all relevant provisions. F1 Score is the harmonic mean of Precision and Recall.

5.3 Experimental Results

Table 2 presents the overall performance of our proposed approach and data across several models, including BM25, BiEncoder Ranker, RAG, and voting with RAG. Specifically, the Ranker refers to the model discussed in Section 4.2 and we compare LLaMA [14] and TaiwanLLaMA [2, 8] for RAG.

For the BM25 baseline, we observe a remarkable improvement of 94.05% in F1-score performance with the style-translated data, highlighting the significant impact of our proposed data transformation. For the other models, all results using the style-translated data outperform those with the original data, further emphasizing the effectiveness of the style-translated data in enhancing model performance. Additionally, we observe notable improvements in the performance of both LLaMA and TaiwanLLaMA models when incorporating the majority voting strategy. This suggests that our voting strategy, which selects provisions that appear at least m times from the n generated results, helps mitigate randomness and output inconsistencies in the one-iteration RAG process.

6 DISCUSSION

6.1 Does Using another LLM as the Style Translator Work?

With the growing availability of LLMs, it is important to assess whether different models can effectively serve as style translators. To investigate this, we conducted an additional experiment using `Gemini-1.5-Flash` as the style translator. As shown in Table 3, four out of five models achieved higher F1-scores when using the translated data. This result confirms that employing different LLMs for style translation is effective, demonstrating that the approach is not limited to a specific model.

6.2 Does both Query and Provisions Need to be Style Translated?

While applying style translation to both queries and provisions ensures consistency, it also introduces significant computational overhead. To evaluate whether translating both components is necessary, we conducted experiments using different combinations of original and style-translated data. As shown in Table 4, translating either the query or the provision improves accuracy, but the most substantial performance gains occur when both are translated. This finding underscores the importance of style alignment in LQPR and demonstrates that applying style translation to both queries and provisions yields the best results.

Table 3: Performance of Using Another LLM as Style Translator (F1 Score)

Models	Original	Style Translated
BM25	0.0604	**0.1053**
Ranker	0.1750	**0.1988**
LLaMA	0.1408	**0.1590**
LLaMA_voting	0.1633	**0.1790**
TaiwanLLaMA	0.1808	**0.1837**
TaiwanLLaMA_voting	**0.2057**	0.1963

Table 4: Performance of Different Translation Setting of Query and Provisions

Query / Provision	Original	Translated
Original	0.20569	0.21248
Translated	0.20862	**0.21876**

Table 5: Performance of direct fine-tuning LLM (F1-score)

Models	Original	Style Translated
LLaMA	0.1601	**0.1618**
TaiwanLLaMA	0.1891	**0.2211**

6.3 Can direct fine-tuning also benefit from style translation?

Fine-tuning LLMs has become a prevalent approach in the question-answering domain. In this subsection, we investigate whether style translation enhances LQPR performance in fine-tuned models. We conducted experiments using zero-shot prompt on original data and translated data. For fine-tuning, we trained Taiwan-LLaMA and LLaMA using QLoRA for 200 steps, and the results are presented in Table 5. The results indicate that, for both models, style-translated data consistently outperforms the original data. This demonstrates that style translation enhances performance, aligning with our design objectives and further reinforcing its effectiveness in LQPR.

7 CONCLUSION AND FUTURE WORK

In this study, we address the challenge of Legal Query-to-Provision Retrieval, where significant writing style differences between user queries and legal provisions hinder effective retrieval. To bridge this gap, we construct a Query2Provision dataset and introduce a style translation approach that reformulates queries and provisions into more compatible forms. Experimental results demonstrate that style transformation significantly boosts retrieval accuracy, highlighting its importance in legal Retrieval tasks. Exploring a wider range of prompts and datasets presents a valuable direction for future research. Our findings provide valuable insights into improving legal information retrieval, and our dataset and methods serve as a foundation for future research in legal AI, question answering, and document retrieval systems.

Acknowledgments

This work was supported by National Science and Technology Council, Taiwan, under grants NSTC 113-2634-F-002-003-, and Ministry of Education (MOE), Taiwan, under grants NTU-113L900901.

References

[1] Andong Chen, Feng Yao, Xinyan Zhao, Yating Zhang, Changlong Sun, Yun Liu, and Weixing Shen. 2023. EQUALS: A Real-world Dataset for Legal Question Answering via Reading Chinese Laws. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law* (2023). https://api.semanticscholar.org/CorpusID:261583507

[2] Po-Heng Chen, Sijia Cheng, Wei-Lin Chen, Yen-Ting Lin, and Yun-Nung Chen. 2024. Measuring Taiwanese Mandarin Language Understanding. In *First Conference on Language Modeling*. https://openreview.net/forum?id=7jSMMvXLri

[3] Zhe Chen, Fuhui Sun Pengjie Ren, Xiaoyan Wang, Yujun Li, Siwen Zhao, and Tengyi Yang. 2025. SLARD: A Chinese Superior Legal Article Retrieval Dataset. In *Proceedings of the 31st International Conference on Computational Linguistics*. 740–754. https://aclanthology.org/2025.coling-main.50/

[4] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-Training With Whole Word Masking for Chinese BERT. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 29 (Nov. 2021), 3504–3514. doi:10.1109/TASLP.2021.3124365

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLORA: efficient finetuning of quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 441, 28 pages.

[6] Yi Feng, Chuanyi Li, and Vincent Ng. 2024. Legal Case Retrieval: A Survey of the State of the Art. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:271911742

[7] Haitao Li, Weihang Su, Changyue Wang, Yueyue Wu, Qingyao Ai, and Yiqun Liu. 2023. Thuir@ coliee 2023: Incorporating structural knowledge into pre-trained language models for legal case retrieval. *arXiv preprint arXiv:2305.06812* (2023).

[8] Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan LLM: Bridging the Linguistic Divide with a Culturally Aligned Language Model. *ArXiv* abs/2311.17487 (2023). https://api.semanticscholar.org/CorpusID:265498587

[9] Antoine Louis and Gerasimos Spanakis. 2022. A Statutory Article Retrieval Dataset in French. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 6789–6803. doi:10.18653/v1/2022.acl-long.468

[10] Antoine Louis, G. van Dijck, and Gerasimos Spanakis. 2023. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:263310713

[11] Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, M. Zhang, and Shaoping Ma. 2021. LeCaRD: A Legal Case Retrieval Dataset for Chinese Law System. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021). https://api.semanticscholar.org/CorpusID:235792264

[12] Francesco Sovrano, Monica Palmirani, Biagio Distefano, Salvatore Sapienza, and Fabio Vitali. 2021. A dataset for evaluating legal question answering on private international law. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (2021). https://api.semanticscholar.org/CorpusID:236459372

[13] Weihang Su, Yiran Hu, Anzhe Xie, Qingyao Ai, Zibing Que, Ning Zheng, Yun Liu, Weixing Shen, and Yiqun Liu. 2024. STARD: A Chinese Statute Retrieval Dataset Derived from Real-life Queries by Non-professionals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 10658–10671. https://aclanthology.org/2024.findings-emnlp.625/

[14] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv* abs/2302.13971 (2023). https://api.semanticscholar.org/CorpusID:257219404

[15] Weiqi Zhang, Hechuan Shen, Tianyi Lei, Qian Wang, Dezhong Peng, and Xu Wang. 2023. GLQA: A Generation-based Method for Legal Question Answering. *2023 International Joint Conference on Neural Networks (IJCNN)* (2023), 1–8. https://api.semanticscholar.org/CorpusID:260387152