# Enhancing Legal Case Retrieval via Scaling High-quality Asymmetric Query-Candidate Pairs

Anonymous ACL submission


## Abstract

Legal case retrieval (LCR) aims to provide similar cases as references for a given fact description. This task is crucial for promoting consistent judgments in similar cases, effectively enhancing judicial fairness and improving work efficiency for judges. However, existing works face two main challenges for real-world applications: existing works mainly focus on case-to-case retrieval using lengthy queries, which does not match real-world scenarios; and the limited data scale, with current datasets containing only hundreds of queries, is insufficient to satisfy the training requirements of existing data-hungry neural models. To address these issues, we introduce an automated method to construct asymmetrically query-candidate pairs and construct the largest LCR dataset to date, LEAD, which is hundreds of times larger than existing datasets. This dataset can provide ample training signals for LCR models. Experimental results demonstrate that models training with LEAD can achieve state-of-the-art results on two widely-used LCR benchmarks. Besides, the construction method can be also applied to civil cases and achieve promising results. The code and dataset used in this paper will be released to promote the development of LCR.


## 1 Introduction

Legal case retrieval (LCR) aims to search for historically relevant cases based on a given fact description (Bench-Capon et al., 2012; Bhattacharya et al., 2022; Locke and Zuccon, 2022; Yu et al., 2022; Sansone and Sperlí, 2022). This task can help legal professionals, such as judges and lawyers, improve work efficiency by providing past cases as references for current judgments. Thus, it plays a crucial role in promoting judicial fairness by facilitating similar cases receiving similar judgments.

Different from open-domain retrieval, LCR demands a complex understanding of case details and necessitates models equipped with legal knowledge to generate knowledge-rich case representations (Xiao et al., 2023; Sun et al., 2023a). As shown in Figure 1, models are required to recognize that the severity of injury rather than the location of injury is the key factor in assessing the relevance of given candidates to the query. Recent years have seen significant efforts by scholars to improve the performance of LCR, including introducing additional knowledge features (Bhattacharya et al., 2022; Yao et al., 2022; Sun et al., 2023a) and designing LCR-oriented pre-training objectives (Li et al., 2023a; Ma et al., 2023).

**Query:** Someone **(1) injured** another person, causing multiple injuries to the head and chest, which were assessed as **(2) minor and moderate injuries**.

**Candidate Case 1:** … During the fight, Bob **(1) punched** Charlie, causing a fracture to the lower section of his right ulna bone… Charlie's injuries were classified as **(2) moderate injuries**…

**Candidate Case 2:** … During the fight, Bob **(1) stabbed** Charlie in the head and chest… Charlie's injuries were classified as **(2) severe injuries** …

**Relevance: Case 1 > Case 2**

Figure 1: An example for legal case retrieval, where the key facts are in blue.

However, despite these advancements, the real-world application of LCR still faces the following challenges: (1) **Asymmetric Retrieval.** Existing methods mostly focus on symmetric retrieval settings with lengthy fact descriptions for both queries and candidates. In contrast, real-world user queries often consist of only a few sentences describing key details. This inconsistency between application and training scenarios results in sub-optimal performance. (2) **Limited Data.** Another challenge is the limited data scale, as legal data annotation requires highly skilled and experienced annotators, making it time-consuming and labor-intensive. Existing LCR datasets contain only a few hundred

queries (Ma et al., 2021; Li et al., 2023b), compared to tens of thousands in open-domain retrieval datasets (Bonifacio et al., 2021; Qiu et al., 2022; Xie et al., 2023). Besides, most retrieval methods rely heavily on data-hungry neural models, making the construction of large-scale, high-quality legal retrieval data a key to enhancing LCR performance.

To address these issues, this paper proposes a method for automatically constructing a high-quality, asymmetric legal retrieval dataset, LEAD. Specifically, given a case candidate, we employ a large-scale generative language model to first extract key facts, and omit entities, including names and places. Then, based on the anonymous key fact, we require the model to generate a brief and coherent description of the case, which is regarded as the search query. In this way, the generated query is short and contains only a few sentences. Additionally, to improve the diversity of LEAD and enable the model to retrieve relevant cases even when key facts are not entirely consistent, we employ a knowledge-driven data augmentation strategy. For each query, we select the case that is most similar from the perspective of charges, related legal articles, and prison term, from the entire corpus as the augmented positive candidate. This approach enables us to build the largest LCR dataset to date, with over 100K query-candidate pairs, surpassing existing LCR datasets by a hundredfold.

To verify the effectiveness of our method, we train dense passage retrieval models with LEAD and compare the model with several competitive baseline models, on two widely-used criminal LCR benchmarks. The experimental results demonstrate that models trained with our enriched high-quality case retrieval data can achieve state-of-the-art performance in LCR tasks. Besides, the proposed framework for data generation can be easily applied to civil case retrieval, and achieve satisfying performance. The code and data in our paper will be released to promote the development of LCR.

## 2 Related Work

**Legal Case Retrieval** Legal case retrieval is a challenging task that requires a deep understanding of legal documents. The task entails models identifying the most legally relevant cases within candidate documents concerning a given query case.

Earliest work for LCR attempt to employ traditional retrieval models, including, BM25 (Robertson and Zaragoza, 2009) and TF-IDF (Aizawa, 2003), for legal retrieval (Zeng et al., 2007). With the development of deep learning, many efforts have been devoted into designing neural architectures to enhance long textual representation (Beltagy et al., 2020; Shao et al., 2020), interpretability (Yu et al., 2022; Sun et al., 2023b), legal knowledge enriched representation (Abolghasemi et al., 2022; Ma et al., 2024; Xiao et al., 2023; Sun et al., 2022; Yao et al., 2022). Due to the lack of a large-scale LCR dataset, these researches mainly focus on the re-ranking phrase, overlooking the significance of dense passage retrieval (DPR) for high recall rate (Karpukhin et al., 2020). To elevate the data scarcity issues, some researchers explore the self-supervised pre-training for legal DPR. For instance, SAILER (Li et al., 2023a) adopts an asymmetric encoder-decoder architecture, integrating various pre-training objectives to encode rich semantic information across tasks. CaseEncoder (Ma et al., 2023) leverages fine-grained legal provisions to select relevant and irrelevant cases for each query, thus improving the quality of training data. In this paper, we find that our LEAD can further facilitate the LCR performance by scaling the high-quality instances for LCR.

**Dataset for LCR** High-quality data lies in the core of existing data-hungry neural models for LCR. However, due to the highly skilled and experienced annotators required for legal data annotation, existing LCR datasets only contain a few hundred queries. For example, LeCaRD (Ma et al., 2021) consists of a total of 107 queries, each with 100 candidate documents, but only 30 of these documents have been manually annotated for relevance. LeCaRDv2 (Li et al., 2023b) contains 800 queries, with only 30 documents per query annotated for relevance. CAIL2022-LCR is the competition dataset of the Challenge of AI in Law (CAIL) [1]. Compared to these datasets, open-domain retrieval datasets have hundreds of times more queries, such as $T^2$Ranking (Xie et al., 2023) with 307k queries, $\text{DuReader}_{\text{retrieval}}$ (Qiu et al., 2022) with 97k queries, and mMarco-Chinese (Bonifacio et al., 2021) with 516k queries. The lack of large-scale data hinders the development of LCR.

**Data Augmentation for Information Retrieval** Data augmentation aims to increase the amount of training data by heuristically generating new data instances based on existing data. In the context

[1] http://cail.cipsc.org.cn/

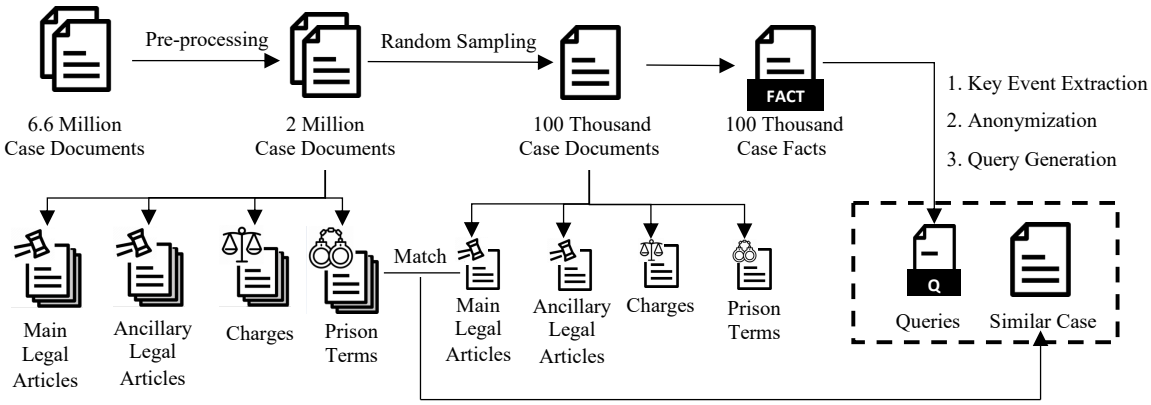


Figure 2: The illustration of the data construction process of LEAD.

of information retrieval, data augmentation is typically applied to generate new queries, positive and negative examples. For example, the Inverse Cloze Task (ICT) (Lee et al., 2019) randomly selects a token span from a text segment to serve as the query, while the remaining tokens form the key. This is the opposite of the Cloze Task, where the remaining tokens are used as the query and the sampled token span serves as the candidate. This approach has been proven effective in pre-training(Chang et al., 2020; Sachan et al., 2021).

Additionally, the use of in-batch negatives is a method to expand negative examples. For a given query, the negatives are generated from the positive examples of other queries within the same batch. This method typically requires a larger batch size to generate more negatives for a query (Chen et al., 2020) and has been widely applied in open-domain retrieval scenarios (Lee et al., 2019; Karpukhin et al., 2020; Izacard et al., 2022).

## 3 Data Construction

To address the challenges of asymmetric retrieval, queries in the training dataset should align with real-world user queries, which are often characterized by brevity and conciseness. As shown in Figure 2, we propose an automatic method to generate queries based on case facts. We will introduce the details about the data generation in this section.

### 3.1 Query Generation

**Key Events Extraction** As all case documents are manually written by judges, there are many details and viewpoints contained in these documents, such as the names of every participant, their relationships, and the court discussion about each event. However, in real life, considering users' unfamiliarity with legal knowledge, the queries they search often only include key factual events. To get the short queries as real-world user queries, we extract key information from the facts of legal cases gathered from online sources. Then, to do this efficiently, automatically, and at a large scale, our approach leverages a generative method based on open-source, large-scale language models. We employ an LLM to generate queries for our dataset. During the generation process, the model is first required to compress provided case facts into concise case descriptions, which only retain essential legal events. To guide the model, we furnish it with a task description and two illustrative examples within the prompt, ensuring effective and accurate query generation. The specific prompt is provided in the Appendix.

**Anonymization** The queries generated using the above method still contain a considerable number of entities such as personal names, locations, and dates. These entities are usually irrelevant to the key events and do not affect the final judgment. Besides, the shared entities between queries and candidates would provide a shortcut to the models, leading models trained on this data assign high relevance scores to the queries and candidates with the same entities and overlook critical legal events. Therefore, we implement a strategy to anonymize these entities. Specifically, we utilize DeepTHULAC [2] for part-of-speech tagging of queries. Subsequently, specific information such as personal names, company names, locations, and time within the queries are replaced with semantically equivalent content. For instance, personal names are replaced with names like "Alice" and "Bob". This

[2]https://github.com/thunlp/DeepTHULAC

| Dataset | LeCaRD | CAIL2022-LCR | COLIEE2021 | COLIEE2022 | LEAD |
|---|---|---|---|---|---|
| Asymmetric | ✖ | ✖ | ✖ | ✖ | ✔ |
| # Query | 107 | 40 | 900 | 1,198 | 100,060 |
| Language | Chinese | Chinese | English | English | Chinese |
| # Charge | 20 | 19 | – | – | 210 |
| Query Length | 445 | 422 | 2,060 | 2,168 | 79 |

Table 1: Details of statistics of existing LCR datasets. The COLIEE dataset does not annotate the corresponding charges for the cases, so this table does not provide such information.

approach enables the model to better grasp the relationships between queries and key information, thereby enhancing the effectiveness of retrieval.

With the key events extraction and anonymization, we can generate a relevant query for every candidate case. The query-candidate pairs can serve as the training signals for LCR models.

### 3.2 Knowledge-Driven Augmentation

Through the aforementioned method, we can construct large-scale query-candidate pairs that contain the same key facts. However, in real applications, we usually cannot find cases that are completely identical to the query. Therefore, to enable the model to handle a diverse range of queries in real-world scenarios, we further propose a knowledge-driven data augmentation method.

Unlike open-domain information retrieval, in the LCR domain, it is not appropriate to judge whether two cases are similar based solely on the factual details of the case. The legal articles applicable to the case and the judgment results are also important (Li et al., 2023c). Therefore, for a given query-candidate pair, we select the cases with similar legal articles and prison terms to the candidate as the augmented positive candidate. Specifically, we extract the main and ancillary legal articles from the "Reason" section of the case. Here, the main legal articles refer to those detailing specific charges, such as *Article 133* from the Chinese Criminal Law, which defines and sets sentencing standards for the crime of traffic accidents. The ancillary legal articles refer to those outlining the impact of certain facts on sentencing, such as *Article 67* from the Chinese Criminal Law, which defines self-surrender and its influence on the final sentencing. Additionally, we extract the charges and specific prison terms of the final judgment, such as death penalty and imprisonment, from the "Judgment" section. These extracted elements serve as the basis for positive augmentation.

Next, for each candidate case in the dataset, we identify a related case in which the main legal articles match those of the original candidate case, and the additional legal articles as well as prison terms are as similar as possible. This process results in a new positive example. This positive example is legally related to the original case, but because they are two completely different cases, it ensures that there is no overlap in the factual details. This process leads to a dataset that has been augmented with positive examples.

### 3.3 Construction Details

We collect 6.6 million criminal cases from China Judgment Online [3]. Initially, we exclude criminal ruling documents (containing only content related to commutation) and retain only criminal judgment documents. Subsequently, we filter out cases with facts shorter than 100 Chinese characters, as the majority of criminal cases fall within this range. Using regular expressions, we match and extract information such as charges, legal articles, and judgments from the cases, eliminating those where such content couldn't be extracted via rules. In the end, there are about 2 million cases remained. From this pool, we randomly select 100 thousand cases to generate queries for each charge. Then, for each of these 100 thousand cases, we search for the most similar cases from the initial 2 million using charges, legal articles, and judgments as criteria, to augment new positive examples.

### 3.4 Data Analysis

We present the statistics of our LEAD and other widely-used LCR datasets in Table 1. From the results, we can observe that the LEAD dataset is currently the largest LCR dataset, several hundred times larger than the largest datasets available, and capable of supporting the training of existing data-hungry dense passage retrieval models. Furthermore, LEAD is currently the only dataset designed for asymmetric retrieval, which can better serve real-world scenarios. Due to the flexibility of our construction method, it can be extended to any

[3] https://wenshu.court.gov.cn/

charge and case, hence LEAD covers the most extensive range of case charges.

### 3.5 Model Training

In this paper, we mainly focus on dense passage retrieval for legal cases. We adopt a dual-encoder architecture for all models. This involves separately encoding the query and the candidate cases to obtain query embeddings and candidate case embeddings and calculating the cosine similarity between them as the final similarity score.

For model training, we employ a pre-trained language model, Lawformer (Xiao et al., 2021), as the backbone model. The training is conducted in an in-batch negative setting (Karpukhin et al., 2020). In the in-batch negative setting, for each query in a batch with N training pairs, the negative examples are the positives of the other queries in the same batch, i.e., N-1 negative examples. However, when we use the newly identified positive examples from the dataset, some negatives may share the same charges, legal articles, or judgments with the positives, leading to false negatives that can impact the model training. To address this, during training, we straightforwardly set the cosine similarity between negatives with the same charges as the positive to $-\infty$. This is equivalent to removing these negatives from the negative set.

## 4 Experiments

### 4.1 Datasets and Metrics

In this paper, we focus on legal asymmetric retrieval, but existing datasets with human-annotated labels focus on symmetrical retrieval, where the queries are lengthy cases. Therefore, to better assess the model's performance in asymmetric retrieval, we adopt our method to simplify the query cases into a short version automatically. To ensure the high quality of evaluation benchmarks, we manually check the generated queries, ensuring that the queries do not change the key events. Specifically, we employ GPT-4 to generate the short version of queries and conduct quality testing by one of the authors. For case-to-case retrieval, we utilize the original datasets without query generation.

We adopt LEAD for training, and adopt two widely-used datasets for evaluation: (1) **LeCaRD** (Ma et al., 2021) is a widely-used LCR evaluation dataset, which contains 107 queries annotated by several legal practitioners. (2) **CAIL2022-LCR** [4] official testing set is furnished by the CAIL2022 organization, structured similarly to LeCaRD. We test our models on stage 2 of CAIL2022. In both datasets, each query has 100 candidate cases, but only 30 of them are manually annotated. The annotations range from 0 (Both key facts and key circumstances are irrelevant) to 1 (Key facts are irrelevant but key circumstances are relevant), 2 (Key facts are relevant but key circumstances are irrelevant), and 3 (Both key facts and key circumstances are relevant). During evaluation, we consider only the annotated cases, and we only regard cases marked as 3 as relevant, while the rest are deemed irrelevant.

As a retrieval task, we report normalized discounted cumulative gain (NDCG@10, NDCG@20, NDCG@30), Precision (P@5, P@10), and Mean Average Precision (MAP). These evaluation metrics align with those used in LeCaRD, aiming to provide a comprehensive understanding of the model's performance across various aspects.

### 4.2 Baselines

We compare our model with several competitive baselines, including:

*Traditional Retrieval Model*: (1) **BM25** (Robertson and Zaragoza, 2009) utilizes exact word matching to score documents based on their term frequencies and document lengths.

*Pretrained Models*: (1) **Chinese BERT** is an adaptation of the original BERT model (Devlin et al., 2018) for the Chinese. (2) **Lawformer** (Xiao et al., 2021) is the first Chinese legal pre-trained model based on the longformer model (Beltagy et al., 2020). (3) **SAILER** (Li et al., 2023a) is a structure-aware pre-trained model for LCR, which employs an asymmetric encoder-decoder architecture for pre-training.

*Data Augmentation Method*: (1) **Inverse Cloze task (ICT)** (Lee et al., 2019) is a data augmentation method in open-domain for retriever pre-training, which involves randomly sampling a token span from a text segment as the query, while the remaining tokens as the candidate. (2) **CaseEncoder** (Ma et al., 2023) is a legal document encoder that constructs LCR data with fine-grained legal article information, which assumes that similar cases should contain similar legal articles.

*Fine-Tuned Models*: (1) **T$^2$Ranking** (Xie et al.,

[4] http://cail.cipsc.org.cn/task3.html?raceID=3&cail_tag=2022

| Model | Model Type | LeCaRD | | | | | |
|---|---|---|---|---|---|---|---|
| | | P@5 | P@10 | MAP | NDCG@10 | NDCG@20 | NDCG@30 |
| BM25 | Traditional | 44.8 | 40.8 | 50.7 | 77.3 | 82.0 | 89.9 |
| Chinese BERT | Pre-trained | 36.5 | 34.5 | 41.9 | 70.5 | 77.6 | 86.8 |
| Lawformer | Pre-trained | 40.6 | 38.5 | 45.6 | 74.4 | 80.0 | 88.5 |
| SAILER | Pre-trained | 51.8 | 46.5 | 59.7 | 86.0 | 89.5 | 93.9 |
| ICT | Augmentation | 37.6 | 36.7 | 45.6 | 72.2 | 78.9 | 87.5 |
| CaseEncoder | Augmentation | 50.8 | 45.8 | 57.7 | 83.6 | 87.4 | 92.7 |
| $T^2$Ranking | Fine-tuned | 43.7 | 40.0 | 49.3 | 75.6 | 81.6 | 88.9 |
| Ours | Fine-tuned | **56.3** | **49.6** | **63.5** | **87.3** | **89.9** | **94.5** |

| Model | Type | CAIL2022-LCR | | | | | |
|---|---|---|---|---|---|---|---|
| | | P@5 | P@10 | MAP | NDCG@10 | NDCG@20 | NDCG@30 |
| BM25 | Traditional | 54.0 | 49.7 | 57.6 | 81.8 | 86.0 | 91.8 |
| Chinese BERT | Pre-trained | 45.5 | 45.8 | 50.7 | 74.8 | 80.0 | 88.4 |
| Lawformer | Pre-trained | 53.0 | 50.5 | 57.5 | 84.5 | 87.9 | 93.0 |
| SAILER | Pre-trained | 60.5 | 54.2 | 65.7 | 91.9 | 94.3 | 97.0 |
| ICT | Augmentation | 51.0 | 47.7 | 53.5 | 81.5 | 85.2 | 91.5 |
| CaseEncoder | Augmentation | 58.0 | 54.2 | 63.6 | 91.7 | 93.6 | 96.5 |
| $T^2$Ranking | Fine-tuned | 54.5 | 52.2 | 59.3 | 86.6 | 89.4 | 94.1 |
| LeCaRD Train | Fine-tuned | 56.0 | 53.5 | 59.6 | 88.6 | 91.5 | 94.7 |
| Ours | Fine-tuned | **65.0** | **58.0** | **67.7** | **94.0** | **94.7** | **97.4** |

Table 2: The main results of our model trained on LEAD and baseline models on LeCaRD and CAIL2022-LCR under the asymmetric retrieval setting.

2023) is a large-scale retrieval dataset in the open-domain. We directly utilize an open-source dual-encoder checkpoint, fine-tuned on the $T^2$Ranking dataset as our baseline model. (2) **LeCaRD Train** refers to the models trained with the instances contained in LeCaRD. For a fair comparison, we adopt the Lawformer as the backbone model. Here, as LeCaRD is used for training, we only present the results of the model for CAIL2022-LCR.

### 4.3 Implementation Details

During evaluation, we employ a truncation strategy for lengthy candidates. Specifically, when the length of a candidate case exceeds the maximum sequence length of the utilized models, we truncate the case into multiple segments. Subsequently, we individually calculate the similarity score between each segment and the query, ultimately selecting the maximum similarity score as the final score for the candidate case.

The training batch size is set as 128 and the encoders are trained for up to 80 epochs with a learning rate of 1e-5 using Adam, linear scheduling with warm-up, and dropout rate 0.1. The maximum input sequence length was set to 2048. Additionally, our model reported in Table 2 utilizes positive augmentation data at a ratio of 70%. That is, 30% of the query-candidate pairs in the dataset consist of queries paired with their original cases, while the remaining 70% of query-candidate pairs comprise simplified queries paired with cases newly identified using the method outlined in Section 3.2. We randomly select 2048 samples from the dataset as the development set, with the rest used for training.

### 4.4 Main Result

The overall results are presented in Table 2. From the results, we can observe that: (1) Our model outperforms all baselines on both benchmarks by a large margin, achieving state-of-the-art performance. It indicates that LEAD, consisting of high-quality LCR instances, can greatly benefit the task performance. (2) The traditional method, BM25, can outperform many models. Especially, BM25 can beat the models finetuned on $T^2$Ranking, which consisting millions of open-domain retrieval instances. It proves that LCR task is challenging and directly employing open-domain models can not achieve satisfactory results. That is because LCR requires the models to capture not only semantic relevance but also legal element relevance. (3) Compared to the pre-trained models, our model trained with LEAD can achieve siginificant performance improvements. The pre-training for LCR usually involves millions of cases and days of pre-training, which is computationally expensive. It shows the potential of scaling high-quality data for LCR, which can avoid expensive pre-training and yield superior performance. Furthermore, our dataset can also be employed on pre-trained models

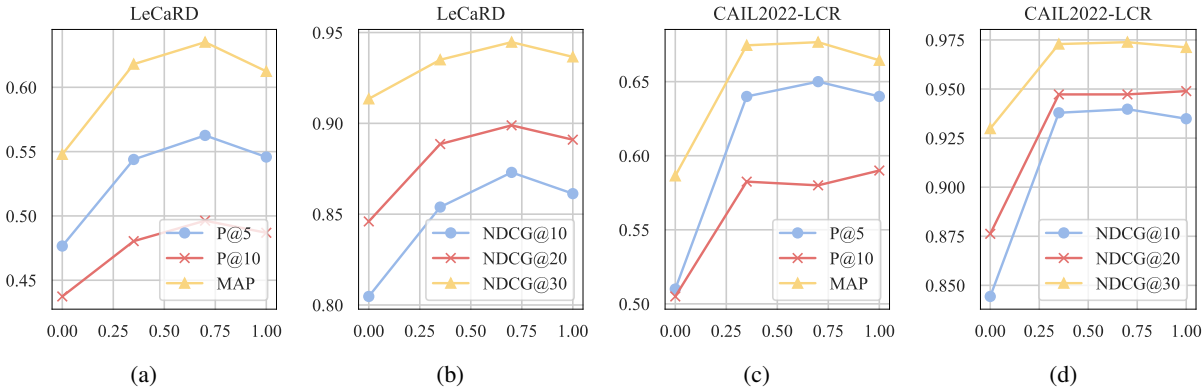


Figure 3: Comparison of model performance with different proportions of augmented positive examples on LeCaRD and CAIL2022-LCR Datasets.

| | LeCaRD | | | |
|---|---|---|---|---|
| | P@5 | MAP | NDCG@10 | NDCG@30 |
| Ours | **56.3** | **63.5** | **87.3** | **94.5** |
| w/o M | 52.0 | 58.0 | 84.1 | 92.8 |
| | **CAIL2022-LCR** | | | |
| | P@5 | MAP | NDCG@10 | NDCG@30 |
| Ours | **65.0** | **67.7** | **94.0** | **97.4** |
| w/o M | 59.5 | 63.4 | 90.4 | 96.1 |

Table 3: Comparison of model performance with and without false negative masking.

| Models | BM25 | BERT | $T^2$Ranking | Ours |
|---|---|---|---|---|
| Accuracy | 54.3 | 52.1 | 52.2 | **56.2** |

Table 4: The results on the CAIL2019-SCM dataset.

such as SAILER for further performance improvements, which we leave for future work. (4) Our model can consistently outperform the data augmentation models and fine-tuned models. The existing data augmentation method can not generate high-quality data for LCR. Besides, existing open-domain data cannot benefit LCR performance, and the scale of existing LCR datasets cannot fulfill the requirements of training dense retrieval models. Our proposed method to automatically construct data is effective in high-quality data generation.

### 4.5 Ablation Study

We adopt a knowledge-driven data augmentation strategy for dataset construction. In this subsection, we conduct an ablation study to explore the impact of augmented positive examples.

**Proportion of Augmented Candidates** We adopt a knowledge-driven data augmentation strategy to make the query-candidate pairs with similar legal elements but diverse legal events. In this paragraph, to verify the effectiveness of the data augmentation, we conduct experiments with varying proportions of augmented positive examples within the dataset. Specifically, we present the results with the proportions as $\{0.00, 0.35, 0.700, 1.00\}$. The results are shown in Figure 3.

From the results, we can observe that: (1) Compared with models without data augmentation (0%), models trained with further data augmentation can achieve significant performance improvements for both two datasets and all metrics. It indicates that the knowledge-driven data augmentation methods can effectively match similar cases from the entire corpus and benefit the diversity of LEAD. (2) The optimal performance is achieved at 70% and when the proportion reaches 100%, the model performance drops. This suggests that retaining a certain proportion of original cases as positive candidates is effective for LCR. We believe this is because these data instances help reduce the distance between simplified queries and original cases in the vector representation space, allowing the model to better comprehend the meaning of simplified queries in asymmetric retrieval scenarios.

**False Negative Masking** We adopt the in-batch negative sampling strategy to increase the scale of negative sampling. However, this training strategy will inevitably introduce false negative noises. To address this challenge, we adopt a false negative masking strategy, where the cosine similarity of negative candidates with the same charges is set to $-\infty$ during the training process. In this paragraph, we evaluate the effects of false negative masking strategy, with the results presented in Table 3. We can find that removing the false negative

| Model | Model Type | CAIL2022-LCR | | | | | |
|---|---|---|---|---|---|---|---|
| | | P@5 | P@10 | MAP | NDCG@10 | NDCG@20 | NDCG@30 |
| BM25 | Traditional | 50.5 | 49.8 | 55.1 | 80.2 | 82.7 | 90.5 |
| Chinese BERT | Pre-trained | 46.5 | 47.0 | 52.6 | 78.2 | 81.8 | 89.9 |
| Lawformer | Pre-trained | 52.0 | 50.8 | 54.9 | 82.6 | 84.6 | 91.2 |
| SAILER | Pre-trained | 60.5 | 55.3 | 66.8 | 92.6 | 94.2 | 97.1 |
| ICT | Augmentation | 48.5 | 47.0 | 52.2 | 79.6 | 82.9 | 90.6 |
| CaseEncoder | Augmentation | 63.5 | 56.0 | 65.6 | 92.8 | 94.1 | 96.9 |
| $T^2$Ranking | Fine-tuned | 56.5 | 50.8 | 57.4 | 83.4 | 86.7 | 92.2 |
| LeCaRD Train | Fine-tuned | 57.0 | 55.6 | 58.6 | 88.1 | 90.9 | 93.8 |
| Ours | Fine-tuned | **65.0** | **58.5** | **69.2** | **94.4** | **95.2** | **97.6** |

Table 5: The results of our model trained on LEAD and baseline models on CAIL2022-LCR under the traditional case-to-case symmetric retrieval setting.

masking strategy significantly deteriorates model performance on both datasets. This suggests that during the training process, many negative examples are indeed related to the query, and ignoring them can mitigate such interference.

### 4.6 Civil Case Retrieval

Our method to automatically construct LCR datasets is flexible and can be easily extended to any case. Existing LCR works usually focus on criminal cases and overlook civil cases, which are more relevant to our daily lives. In this subsection, we construct a civil case retrieval dataset with the same construction method. Specifically, the judgment results of civil cases are more complex than criminal cases, and the knowledge-driven data augmentation strategy cannot be applied to civil cases. Therefore, here we present the results with no further candidate augmentation. Finally, we generate 77k query-candidate pairs for civil cases. We utilize CAIL2019-SCM (Xiao et al., 2019) as the benchmark, which comprises 3036 triplets for the private lending cases, each consisting of three cases' fact descriptions: A, B, and C. The task is to determine which of the descriptions, B or C, is more similar to A. We report the accuracy of several models that are not limited to criminal cases, and our model on this test set in Table 4. Despite using only simplified queries and their corresponding original cases as training data, our model can achieve the best performance on this test set. This demonstrates that simple asymmetric retrieval data can also enable the model to understand legal elements, validating the robustness of our approach.

### 4.7 Case-to-Case Symmetric Retrieval

In this paper, we mainly focus on asymmetric LCR and our large-scale dataset can also benefit the traditional case-to-case symmetric retrieval setting. In this subsection, we evaluate the models in the traditional setting. The results are shown in Table 5. From the results, we can observe that (1) Our model still outperforms other models by a large margin, indicating that our constructed asymmetric retrieval dataset, LEAD, is not only effective for asymmetric retrieval tasks but also performs excellently in traditional case retrieval scenarios. This suggests that our model effectively learns to identify similar legal elements through augmented positive examples. (2) The baseline models can achieve superior performance on the asymmetric retrieval setting. That is because the lengthy query can provide more detailed information for models to retrieve similar cases. The short queries require the models to associate the key events and legal knowledge to capture relevance between the query and candidates, which presents a great challenge for existing models. Therefore, we encourage the community to devote more efforts to asymmetric LCR for real-world applications.

## 5 Conclusion

In this paper, we propose a method for automatically constructing high-quality, asymmetric legal case retrieval datasets. Our approach leverages a large-scale generative language model to generate simplified queries based on case facts, we further enhance positive examples with a knowledge-driven data augmentation method. We construct the largest legal case retrieval dataset to date, with over one hundred thousand query-candidate pairs, surpassing existing datasets by a hundredfold. We conduct experiments on two widely-used datasets, achieving state-of-the-art performance in legal case retrieval tasks, with a notably significant margin. Moreover, our method is highly versatile, showing superior performance in civil case retrieval as well.

## Limitations

In this paper, we discuss the limitations of this paper: (1) We construct a large-scale asymmetric LCR dataset for Chinese cases. Our method is language-agnostic and can also be applied to cases in other countries, which is worth exploring in the future. (2) In this paper, we conduct experiments with Lawformer (Xiao et al., 2021) with 110M parameters as the backbone. Exploring LCR with larger pre-trained models can further improve the performance.

## References


Amin Abolghasemi, Suzan Verberne, and Leif Azzopardi. 2022. Improving bert-based query-by-document retrieval with multi-task optimization. In *Proceedings of ECIR*, volume 13186 of *Lecture Notes in Computer Science*, pages 3–12. Springer.

Akiko N. Aizawa. 2003. An information-theoretic perspective of tf-idf measures. *Inf. Process. Manag.*, 39(1):45–65.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Trevor J. M. Bench-Capon, Michal Araszkiewicz, Kevin D. Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Danièle Bourcier, Paul Bourgine, Jack G. Conrad, Enrico Francesconi, Thomas F. Gordon, Guido Governatori, Jochen L. Leidner, David D. Lewis, Ronald Prescott Loui, L. Thorne McCarty, Henry Prakken, Frank Schilder, Erich Schweighofer, Paul Thompson, Alex Tyrrell, Bart Verheij, Douglas N. Walton, and Adam Z. Wyner. 2012. A history of AI and law in 50 papers: 25 years of the international conference on AI and law. *Artif. Intell. Law*, 20(3):215–319.

Paheli Bhattacharya, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2022. Legal case document similarity: You need both network and text. *IPM*, 59(6):103069.

Luiz Henrique Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Frassetto Nogueira. 2021. mmarco: A multilingual version of MS MARCO passage ranking dataset. *CoRR*, abs/2108.13897.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *Proceedings of ICLR*. OpenReview.net.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *TMLR*, 2022.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, pages 6769–6781. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of ACL*, pages 6086–6096. Association for Computational Linguistics.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023a. SAILER: structure-aware pre-trained language model for legal case retrieval. In *Proceedings of SIGIR*, pages 1035–1044. ACM.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2023b. Lecardv2: A large-scale chinese legal case retrieval dataset. *CoRR*, abs/2310.17609.

Qingquan Li, Yiran Hu, Feng Yao, Chaojun Xiao, Zhiyuan Liu, Maosong Sun, and Weixing Shen. 2023c. MUSER: A multi-view similar case retrieval dataset. In *Proceedings of CIKM*, pages 5336–5340. ACM.

Daniel Locke and Guido Zuccon. 2022. Case law retrieval: problems, methods, challenges and evaluations in the last 20 years. *CoRR*, abs/2202.07209.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: A legal case retrieval dataset for chinese law system. In *Proceedings of SIGIR*, pages 2342–2348. ACM.

Yixiao Ma, Yueyue Wu, Qingyao Ai, Yiqun Liu, Yunqiu Shao, Min Zhang, and Shaoping Ma. 2024. Incorporating structural information into legal case retrieval. *ACM Trans. Inf. Syst.*, 42(2):40:1–40:28.

Yixiao Ma, Yueyue Wu, Weihang Su, Qingyao Ai, and Yiqun Liu. 2023. Caseencoder: A knowledge-enhanced pre-trained model for legal case encoding. In *Proceedings of EMNLP*, pages 7134–7143. Association for Computational Linguistics.

Yifu Qiu, Hongyu Li, Yingqi Qu, Ying Chen, Qiaoqiao She, Jing Liu, Hua Wu, and Haifeng Wang. 2022. Dureader-retrieval: A large-scale chinese benchmark

for passage retrieval from web search engine. In *Proceedings of EMNLP*, pages 5326–5338. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Devendra Singh Sachan, Siva Reddy, William L. Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. In *Proceedings of NeurIPS*, pages 25968–25981.

Carlo Sansone and Giancarlo Sperlí. 2022. Legal information retrieval systems: State-of-the-art and open issues. *Inf. Syst.*, 106:101967.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. BERT-PLI: modeling paragraph-level interactions for legal case retrieval. In *Proceedings of IJCAI*, pages 3501–3507. ijcai.org.

Zhongxiang Sun, Jun Xu, Xiao Zhang, Zhenhua Dong, and Ji-Rong Wen. 2022. Law article-enhanced legal case matching: a model-agnostic causal learning approach. *CoRR*, abs/2210.11012.

Zhongxiang Sun, Jun Xu, Xiao Zhang, Zhenhua Dong, and Ji-Rong Wen. 2023a. Law article-enhanced legal case matching: A causal learning approach. In *Proceedings of SIGIR*, pages 1549–1558. ACM.

Zhongxiang Sun, Weijie Yu, Zihua Si, Jun Xu, Zhenhua Dong, Xu Chen, Hongteng Xu, and Ji-Rong Wen. 2023b. Explainable legal case matching via graph optimal transport. *TKDE*.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Chaojun Xiao, Zhiyuan Liu, Yankai Lin, and Maosong Sun. 2023. Legal knowledge representation learning. In *Representation Learning for Natural Language Processing*, pages 401–432. Springer Nature Singapore Singapore.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Heng Wang, Jianfeng Xu, et al. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv preprint arXiv:1911.08962*.

Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, and Jin Ma. 2023. T2ranking: A large-scale chinese benchmark for passage ranking. In *Proceedings of SIGIR*, pages 2681–2690. ACM.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A large-scale chinese legal event detection dataset. In *Findings of ACL*, pages 183–201. Association for Computational Linguistics.

Weijie Yu, Zhongxiang Sun, Jun Xu, Zhenhua Dong, Xu Chen, Hongteng Xu, and Ji-Rong Wen. 2022. Explainable legal case matching via inverse optimal transport-based rationale extraction. In *Proceedings of SIGIR*, pages 657–668. ACM.

Yiming Zeng, Ruili Wang, John Zeleznikow, and Elizabeth A. Kemp. 2007. A knowledge representation model for the intelligent retrieval of legal cases. *Int. J. Law Inf. Technol.*, 15(3):299–319.

# A Appendix

## A.1 Data Construction Details

To generate concise case descriptions from case facts, we employ a large-scale generative language model, for query generation. The input instructions and a sample case description, along with its original case fact, are shown in Table 6.

The generated case description retains all the legal elements from the original case fact while omitting the rest of the content. The original case fact, being part of a court judgment, contains a plethora of details to comprehensively describe the case's proceedings. However, including these details as part of a real-world user query is redundant.

## A.2 Experimental Details

**Training with LeCaRD** LeCaRD training set annotates 30 cases for relevance to each query. When constructing the dataset, for each query $Q_i$, all cases with a relevance score of 3 are designated as $\{P_{i1}, P_{i2}, ..., P_{in}\}$, while the remaining cases are designated as $\{N_{i1}, N_{i2}, ..., N_{im}\}$. If $m < n$, then $m - n$ cases are randomly selected from the 70 unannotated cases to form $\{N_{i(m+1)}, N_{i(m+2)}, ..., N_{in}\}$. Each training datum consists of one query, one positive case, and one negative case, denoted as $(Q_i, P_{ij}, N_{ij})$, where $i = 1, 2, ..., 107$ and $j = 1, 2, ..., n$. This process results in a training set of size 1,112. The remaining implementation details are same as those described in Section 4.3. Existing datasets usually contain limited annotated pairs and cannot fulfill the requirements for the training of data-hungry neural models.

## A.3 Addition Experiment Result

We also conducted experiments on the original LeCaRD dataset under the traditional case-to-case symmetric retrieval setting, and the results are shown in Table 7. Here, we present the results of all baseline models and the models trained on LEAD with different proportions of augmented positive examples.

From the results, we can observe that similar to the results on the CAIL2022-LCR dataset, our dataset, LEAD can significantly benefit the performance of traditional case-to-case symmetric retrieval.

**System:**
As a legal expert, you are capable of extracting key elements from the basic information of a case.
**User:**
I have a dataset for searching cases by case. However, the basic information of the cases in my dataset is too long. I will send you the basic information of these cases, please help me simplify them, and greatly shorten their length while retaining key legal elements. You can remove non-key names, locations, etc., but do not delete important elements for case judgments.
**Assistant:**
I understand your requirement.
**User:**
In August 2015, XX Co., Ltd. obtained the right to use state-owned construction land in XX Lake area of XX District through public auction, and developed the "XX" project. The defendants Alice, Bob, Charlie, David, and Edward are villagers of XX Village, XX Street Office, XX District, within the scope of the project. The demolition and compensation within the scope of the land has been implemented by the Management Committee of XX in 2014. After the "XX" project started construction on July 13, 2016, Alice, Bob, and others proposed to the construction party to contract part of the project such as excavation of retaining walls and earthwork excavation of foundation piles. Because they did not have the construction qualification, the project developer did not agree. From July 20 to August 12, 2016, Alice, Bob, Charlie, David, and Edward convened at "XX" and "XX" Tea House to discuss obstructing the construction, and raised 70,000 yuan for the obstruction fund. At the same time, Bob, Charlie, David, and others used WeChat groups and phone calls to invite and mobilize more than 20 villagers from their village and their family members to obstruct the construction of the "XX" construction site by methods such as locking the gate, pulling the power switch, insulting, and standing on construction machinery, causing the construction site to be unable to proceed normally. The economic loss caused by the obstruction during the construction period was appraised by XX District Price Certification Center as 124,530 yuan.
**Assistant:**
A company obtained the right to use construction land through auction and compensation has been implemented, but villagers nearby conspired to obstruct construction, raising a fund of 70,000 yuan and mobilizing more than 20 villagers to repeatedly obstruct the construction site, causing an economic loss appraised at 124,530 yuan.

Table 6: Prompt and an example for generating short queries, where specific information such as names and locations have been anonymized.

| Model | Model Type | LeCaRD | | | | | |
|---|---|---|---|---|---|---|---|
| | | P@5 | P@10 | MAP | NDCG@10 | NDCG@20 | NDCG@30 |
| BM25 | Traditional | 40.7 | 39.5 | 48.9 | 73.5 | 78.8 | 87.7 |
| Chinese BERT | Pre-trained | 36.8 | 36.0 | 42.8 | 70.2 | 77.0 | 86.5 |
| Lawformer | Pre-trained | 40.2 | 37.7 | 46.7 | 73.6 | 79.7 | 88.3 |
| SAILER | Pre-trained | 49.5 | 44.3 | 57.7 | 84.7 | **88.9** | **93.7** |
| ICT | Augmentation | 36.3 | 35.6 | 45.1 | 70.0 | 77.0 | 86.6 |
| CaseEncoder | Augmentation | 49.2 | 45.8 | 57.2 | 83.5 | 87.5 | 92.9 |
| $T^2$Ranking | Fine-tuned | 43.9 | 40.1 | 49.9 | 75.7 | 81.1 | 89.0 |
| Ours (0%) | Fine-tuned | 45.0 | 42.0 | 51.7 | 77.8 | 82.8 | 90.1 |
| Ours (35%) | Fine-tuned | 51.8 | 46.4 | 59.0 | 83.1 | 87.2 | 92.5 |
| Ours (70%) | Fine-tuned | **54.4** | 47.1 | 60.9 | 84.3 | 87.8 | 93.0 |
| Ours (100%) | Fine-tuned | 52.3 | **47.3** | **61.8** | **84.7** | 88.2 | 93.3 |

Table 7: The results of our model trained on LEAD and baseline models on LeCaRD under the traditional case-to-case symmetric retrieval setting.