LeCaRDv2: A Large-Scale Chinese Legal Case Retrieval Dataset

Haitao Li
DCST, Tsinghua University
Beijing, China
liht22@mails.tsinghua.edu.cn

Yunqiu Shao
DCST, Tsinghua University
Beijing, China
shaoyq18@mails.tsinghua.edu.cn

Yueyue Wu
Quan Cheng Laboratory&
DCST, Tsinghua University
Beijing, China
wuyueyue@mail.tsinghua.edu.cn

Qingyao Ai*
Quan Cheng Laboratory&
DCST, Tsinghua University
Beijing, China
aiqy@tsinghua.edu.cn

Yixiao Ma
DCST, Tsinghua University
Beijing, China
mayx20@mails.tsinghua.edu.cn

Yiqun Liu
Quan Cheng Laboratory&
DCST, Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

As an important component of intelligent legal systems, legal case retrieval plays a critical role in ensuring judicial justice and fairness. However, the development of legal case retrieval technologies in the Chinese legal system is restricted by three problems in existing datasets: limited data size, narrow definitions of legal relevance, and naive candidate pooling strategies used in data sampling.

To alleviate these issues, we introduce LeCaRDv2, a large-scale **Le**gal **Ca**se **R**etrieval **D**ataset (version 2). It consists of 800 queries and 55,192 candidates extracted from 4.3 million criminal case documents. To the best of our knowledge, LeCaRDv2 is one of the largest Chinese legal case retrieval datasets, providing extensive coverage of criminal charges. Additionally, we enrich the existing relevance criteria by considering three key aspects: characterization, penalty, procedure. This comprehensive criteria enriches the dataset and may provides a more holistic perspective. Furthermore, we propose a two-level candidate set pooling strategy that effectively identify potential candidates for each query case. It's important to note that all cases in the dataset have been annotated by multiple legal experts specializing in criminal law. Their expertise ensures the accuracy and reliability of the annotations. We evaluate several state-of-the-art retrieval models at LeCaRDv2, demonstrating that there is still significant room for improvement in legal case retrieval. The details of LeCaRDv2 can be found at the anonymous website https://github.com/THUIR/LeCaRDv2.

CCS CONCEPTS

• **Information systems** → **Retrieval tasks and goals**.

KEYWORDS

legal case retrieval, relevance criteria, candidate pooling

*Corresponding author

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07…$15.00
https://doi.org/10.1145/3626772.3657887

ACM Reference Format:
Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024. LeCaRDv2: A Large-Scale Chinese Legal Case Retrieval Dataset. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626772.3657887

1 INTRODUCTION

As a fundamental component of intelligent legal systems, legal case retrieval technology plays an essential role in ensuring justice in judgments. In countries with case law system, judges need to make a final decision based on the previous judgments of relevant cases [30]. In countries with statutory law system, when a case is presented to the court, extensive relevant cases are reviewed to avoid inappropriate judgments [12]. With the rapid growth of digitized legal cases, more and more researchers have started to look into the problem and try to apply natural language processing (NLP) and information retrieval (IR) techniques to address the problem of legal case retrieval [14, 16, 29].

To facilitate relevant research, researchers have begun to build human-annotated datasets for legal case retrieval in recent years [23, 25, 26, 33]. For instance, Juliano Rabelo et al. [25, 26] provide the Canadian legal dataset COLIEE, which belongs to the case law system. Additionally, Ma et al. [23] introduce new relevance judgment criteria for the Chinese statutory law system and construct LeCaRDv1, a Chinese legal case retrieval dataset comprising queries and corresponding relevant case documents. Despite its valuable contributions to the advancement of legal case retrieval techniques in the Chinese legal system, there are still three primary challenges remaining unsolved:

- **Limited Data.** The LeCaRDv1 only contains a hundred queries with a limited number of annotated case documents, which may not be sufficient for the training of large language models and providing reliable evaluation results. More specifically, LeCaRDv1 has 10,700 candidate cases and 107 query cases covering 20 charges. The small query set size and charge coverage rate could limit the scope of the application of the dataset.
- **Narrow Definition of Legal Relevance.** The relevance criterion of LeCaRDv1 focuses only on the fact description section of a case and ignores the similarities in penalty and procedure.

When it comes to creating high-quality datasets, relevance criterion is a fundamental concern especially in the legal field. In general, the relevance in the legal field differs from generic textual similarity and goes beyond topic relevance [28, 31]. In case law systems datasets like COLIEE, a relevant case is typically defined as a previous case cited by the query case. This criterion may not be applicable in countries with statutory law systems, such as China. As a pioneer, LeCaRDv1 proposes new criteria for guiding experts to determine relevance based on critical factors. However, it only concentrates on the fact section, potentially resulting in partial understanding and biased annotations of result relevance.

- **Naive Candidate Pooling Strategy.** LeCaRDv1 employs three retrieval models, namely TF-IDF [1], BM25 [27], and LMIR [34], to construct a 100-case pool for each query. However, these methods primarily rely on lexical matching and exhibit similar characteristics. Consequently, they may not always provide accurate identification of potential cases for labeling purposes.

To address these challenges, we present LeCaRDv2, a Large-Scale Chinese Legal Case Retrieval Dataset. LeCaRDv2 consists of 800 query cases and 55,192 candidate cases selected from a corpus of over 4.3 million Chinese criminal cases. Compared to LeCaRDv1 with only 20 charges in the query set, LeCaRDv2 has three types of query cases covering 50 charges, which can evaluate the effectiveness of retrieval models in the legal domain more comprehensively. Moreover, with the guidance of official documents published by the Chinese Supreme People's Court, we propose new relevance criteria involving three aspects: characterization, penalty, and procedure. The overall relevance is determined by considering three aspects together. To ensure the quality of the dataset, all annotations are completed by multiple well-trained legal experts who are familiar with all concepts used in the proposed criteria.

Different from previous datasets for legal case retrieval, LeCaRDv2 emphasizes directly retrieving relevant cases from a large legal corpus. This is challenging if we only have a limited budget and use simple sampling strategies as those used in LeCaRDv1. To overcome these limitations and label more potential cases with diverse characteristics, we propose a two-level candidate pooling strategy, which includes a retrieval pooling step and a ranking pooling step. For retrieval pooling, we propose Inverse Provision Frequency (IPF) to measure the similarity of cases based on the law articles. Then, to construct the retrieval pool, we employ three distinct methods with different properties: sparse lexical matching, dense semantic retrieval, and the proposed law article similarity. Each method contributes to the construction of a more comprehensive retrieval pool. Moving to the ranking pooling step, we leverage the runs provided by participants in CAIL2021, which are specifically designed for the ranking of LeCaRDv1 and proven to be successful in the context of criminal law cases. These runs are utilized to rank the cases within the retrieval pool, further refining the selection process.

To analyze characteristics of LeCaRDv2, we implement several state-of-the-art models for evaluation. The experimental results indicate that LeCaRDv2 is challenging, and thus more advanced methods for legal case retrieval should be explored. We believe that LeCaRDv2 is a reliable benchmark, and can encourage more fruitful research in the field. It is important to highlight that our dataset provides relevance among case documents. These case documents are openly accessible on the China Judgment Online [1], a website providing collections of case documents published by Chinese government. All sensitive information in cases has been removed or anonymized in advance by the developer of the website, and we have obtained the rightful licences to release all the corresponding case documents and annotations used in LeCaRDv2. Users can obtain specific content and potential updates based on the provided case titles.

In summary, LeCaRDv2 is highlighted in the following aspects:

(1) LeCaRDv2 contains 55,192 candidate cases and 800 query cases covering 50 charges. To the best of our knowledge, LeCaRDv2 is one of the largest Chinese legal case retrieval datasets with the widest coverage of criminal charges. We believe that it can be a reliable benchmark that promotes relevant research in the field.
(2) Compared to LeCaRDv1, we design more comprehensive relevance criteria guided by the official documents from the Chinese Supreme People's Court. The new criterion takes into account three aspects, including characterization, penalty, and procedure, providing a more holistic perspective on the relevance of the case.
(3) We propose a new two-level candidate pooling strategy to identify potential cases with diverse characteristics. Our strategy consists of a retrieval pooling step and a ranking pooling step. Furthermore, Inverse Provision Frequency (IPF) is proposed to measure the law article similarity of two cases.

The rest of the paper is organized as follows: Section 2 introduces the related work. In Section 3, the process of dataset construction is elaborated. Then, the experimental setting and results are introduced in Section 4. Finally, Section 5 concludes our work and discusses future work.

2 RELATED WORK

We survey related work in terms of datasets, and models of legal case retrieval.

2.1 Datasets

Legal case retrieval is an essential and challenging task for legal intelligence systems. Recently, researchers have constructed various benchmarks to promote progress in relevant research. In this section, we provide a summary of existing legal case retrieval benchmarks.

2.1.1 COLIEE. As a well-known competition in the legal field, the Competition on Legal Information Extraction/Entailment (COLIEE) aims to achieve state-of-the-art methods of information retrieval using legal texts [18, 19, 25, 26]. Specifically, COLIEE focuses on Canadian case law, which requires reading a query case and identifying relevant support cases from the candidate corpus. For example, COLIEE2020 contains 650 query cases and each query has 200 candidates. Participants need to re-rank a limited number of cases per query. Different from the previous dataset, COLIEE2021 does not provide a candidate pool for each query, which means

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

that participants need to find relevant cases from the entire corpus. It consists of 4,415 case files with 950 query cases, of which 650 queries are for training and 250 queries for testing.

Since the COLIEE dataset belongs to the case law system, its relevance is significantly different from those used in the statutory law systems, e.g., Chinese law system. In COLIEE, the cases cited are considered relevant. Nonetheless, Chinese legal case documents lack such citations. As a result, we must develop new relevance criteria that are suitable for the Chinese legal system.

2.1.2 CAIL2019-SCM. To encourage the advancement of the relevant case-matching task, the Chinese AI and Law 2019 Similar Case Matching dataset (CAIL2019-SCM) has been released. CAIL2019-SCM comprises 8,964 triplets, distributed across three legal fields, namely private lending, intellectual property disputes, and maritime affairs. Each triplet contains one query case and two candidate cases, and participants are required to identify which candidate case is more relevant to the query case. However, the task definition of CAIL2019-SCM is substantially different from the actual requirements in practical scenarios, which restricts its application.

2.1.3 LeCaRDv1. LeCaRDv1 is the first legal case retrieval dataset based on the Chinese legal system [23]. It consists of 107 query cases and 10,700 candidate cases selected from a corpus of over 43,000 Chinese legal case documents. To cover queries with varying difficulties and ranges, LeCaRDv1 introduces a novel query sampling strategy that includes both common queries and controversial queries. In terms of relevance, LeCaRDv1 focuses on the basic facts and proposes four-level relevance criteria based on the critical factor. The critical factors consist of key circumstances and key elements. When two cases share comparable critical factors, they are considered to be related. Inspired by LeCaRDv1, we further refine the relevance criteria and enlarge the size of the dataset. We hope that this enhanced resource will make a more substantial contribution to the development of the legal community.

2.2 Models

The development of benchmarks has led to the emergence of models that are tailored for legal case retrieval [8, 15, 17, 29, 32]. Shao et al. [29] have employed a strategy that involves breaking down legal case texts into multiple paragraphs and utilizing BERT to determine the similarity between these paragraphs, which achieves promising ranking performance. Xiao et al. [32] introduce a novel attention model and pre-train a Chinese legal language model that can process thousands of tokens. Paheli Bhattacharya et al. [4] devise a technique that combines both textual and citation network information to estimate the similarity between legal cases, surpassing the performance of methods that rely solely on one type of information. Large neural language models are data-hungry and obtaining the data for legal case retrieval is expensive. To promote the process of legal case retrieval, it is necessary to develop a large-scale and high-quality dataset.

3 DATASET CONSTRUCTION

In this paper, our goal is to construct a large-scale and high-quality dataset for legal case retrieval. The conceptual scheme is illustrated in Figure 1. In the following section, we first describe the task definition of legal case retrieval. Then, we elaborate on the corpus, queries, and relevance criteria of the dataset. Finally, we describe the human annotation process and the data analysis.

3.1 Task Defintion

The task of legal case retrieval is to identify cases related to the query case from the candidate set, which support the decision making process of judges. Specifically, given a query case q and a candidate case set $D = \{d_1, d_2, ..., d_n\}$ where n is the number of candidate cases, the task is to retrieval top-k related cases $D_q^* = \{d_i^* | d_i^* \in D\}$ with the highest relevance between the query q and the document d_i^*.

In the practice of legal case retrieval, queries and candidates are usually long-text documents with complex structures. In general, the query is the fact section of a case document and candidates are entire case documents. Consequently, in this dataset, we construct queries by directly extracting the fact descriptions from the case documents, while omitting other sections such as decision

3.2 Corpus and Preprocessing

To construct this dataset, we collected over 4.3 million criminal case documents from the China Judgment Online. As shown in Figure 2, we divide each case into three sections with regular matching, which consists of fact, reason, and decision. Then, we filter the cases where the fact section is less than 50 characters or involves simple procedures[2].

After data pre-processing, the documents are organized with $(key, value)$ pairs. Moreover, we also extract the charges and the articles of criminal law involved in the case with regular matching. We hope to encourage researchers to explore how to build better retrieval models by utilizing them. It is worth to note that all cases in LeCaRDv2 are publicly available on the China Judgment Online. Users can obtain case details and updates from the website according to the title provided.

3.3 Query Selection

Query sampling is essential for the construction of legal case retrieval datasets. A typical way to construct legal case retrieval queries is to sample cases from the legal corpus randomly and use the fact description section as query text. However, the random sampling of query cases often faces several challenges. On the one hand, the charges divide the cases into different topics, and the number of cases with different charges significantly varies. Random sampling cases to create queries may lead to severe long-tail distribution. On the other hand, users of real legal search systems have different search intents. For example, lawyers and judges may focus more on complicated or controversial queries. The public, which lacks legal knowledge, often wishes to retrieve cases containing as much information as possible. Therefore, it is essential for a high-quality dataset to contain queries with different difficulties and multiple types.

Inspired by LeCaRDv1 [23], we apply a sampling strategy that consists of common query, controversial query and procedural query.

[2]Simple procedure refers to the procedure applied to criminal cases with clear facts, simple circumstances, and minor crimes.

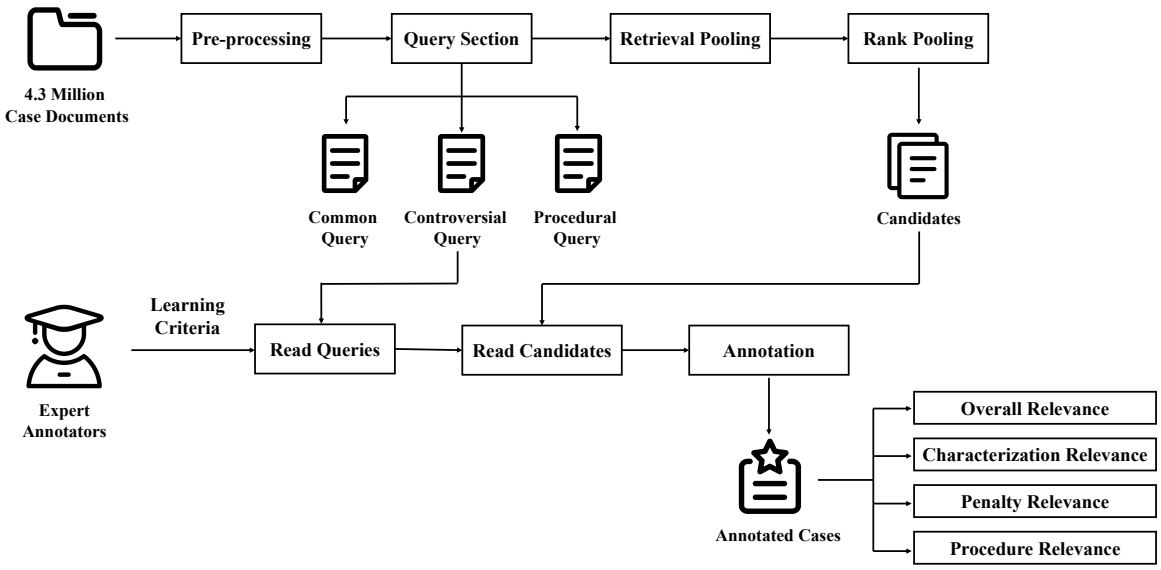


Figure 1: The data collection and annotation process of LeCaRDv2.

Fact: After identification, at 16:25 on August 15, 2017, the defendant drove an overloaded truck and collided with the victim's bicycle riding in the same direction……
Reason: The court held that the defendant violated the traffic regulations and drove a seriously overloaded truck and thus had a major accident……
Decision: According to Article 133 of the Criminal Law of the People's Republic of China, the verdict is as follows: The defendant committed the charge of the traffic accident and was sentenced to one year and five months in prison.

Figure 2: An example legal case document.

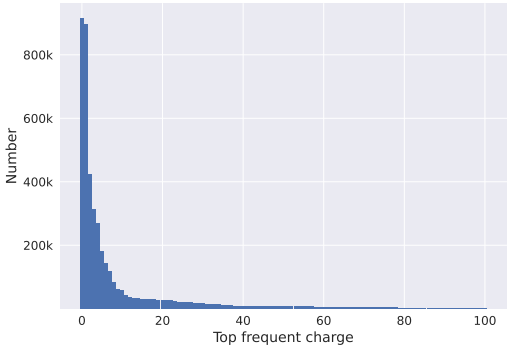


Figure 3: The distribution of top-100 frequent charges of criminal cases in the last 20 years. X-axis shows the charges IDs sorted by their frequency and Y-axis gives the number of the charges.

For common query and controversial query, we expand the coverage of charges used in LeCaRDv1 to include more types of queries. Next, we describe the query sampling strategy in detail.

3.3.1 Common Query. The common query refers to cases without second trials and retrials. In other words, legal experts are more likely to come to a consensus in these cases.

As shown in Figure 5, we compute the statistics of the distribution of charges of criminal cases in the last 20 years. It can be found that the charges of these cases are severely long-tailed distribution. The number of cases with certain charges can vary from several to hundreds of thousands. Models trained on the dataset with a long-tail distribution may have a strong bias in some charges and cannot accurately estimate the relevance of cases with less frequent charges. To overcome this problem, LeCaRDv1 samples queries from top-20 frequent charges evenly, which account for 86.8% of all cases. In this paper, we further expand the coverage of charges and select the top-50 frequent charges to construct the common query. The top-50 frequent charges account for 96.7% of the total number of cases, which can cover most of the query cases.

To satisfy the different search intents of the search system users, a high-quality dataset should contain queries with varying levels of difficulty. Following LeCaRDv1, judgment prediction models are employed to sample queries. Specifically, we first train the judgment prediction model following Zhong et al [36]. Then, we predict the charge of all cases and calculate the prediction entropy to represent the confidence of the prediction model. The prediction entropy can be represented as follows:

$$H(c_i) = -\sum_{j=1}^{N} p_{ij} \log p_{ij} \tag{1}$$

where p_{ij} denotes the probability that the ith case is predicted to be the jth charge and the $H(c_i)$ is the entropy of the ith case. N represents the total number of charges. For a given case, a higher entropy represents a lower confidence in the prediction of the model. Moreover, the charge with the highest probability is considered the final predicted result of the case. According to prediction correctness and prediction entropy, we classify common queries into the following four categories:

- **true-high entropy**: The model predicts the charge correctly but is not confident. This type of query has moderate difficulty.

- **true-low entropy**: The model predicts the correct charge with high confidence. This type of query is usually easy to solve.
- **false-high entropy**: The model predicts the wrong charge but is uncertain. This type of query is difficult.
- **false-low entropy**: The model predicts the wrong charge with high confidence. This type of query is also difficult.

For each charge, we sample three queries in each category. In total, there are $50 \times 12 = 600$ common queries in our dataset. Compared to LeCaRDv1, the number of common queries is extended seven times. This can provide more training signals and more reliable evaluation results.

3.3.2 Controversial Query. The Controversial Query is a case where legal experts have difficulty reaching a consensus. In general, the second trial and retrial cases are more complex and require further discussion. Therefore, we collect all second trial and retrial cases in the corpus to construct the controversial query. Following LeCaRDv1, we sample controversial queries based on the probability of revising judgments.

Specifically, we calculate the probability of the charges change P, the detail of which can be referred to LeCaRDv1 [23]. Then, assuming that the charge C_0 is changed to p charges in total, we arrange them in descending order $P_{C_0 \to C_1}, P_{C_0 \to C_2}, ..., P_{C_0 \to C_p}$. $Top-q$ frequent charges are selected and q satisfies the following conditions:

$$\sum_{i=1}^{q} P_{C_0 \to C_i} \geq 0.5 \tag{2}$$

The controversial queries are sampled from the selected charges. There are a total of 100 controversial queries in this dataset.

3.3.3 Procedural Query. The procedural query refers to cases where procedural legality is disputed in criminal proceedings. An intuitive approach is to select cases related to procedural law. Specifically, substantive law concerns the set of legal principles that govern the behavior of individuals and society generally. Procedural law, also known as adjective law, refers to the regulations and guidelines that govern the processes involved in creating, implementing, and enforcing substantive law. The criteria for relevance under procedural law may vary from those used in substantive law.

To formulate the procedural query, we gather the relevant keywords[3] related to procedural disputes and utilize them to identify a set of appropriate procedural queries. Then, for the top-50 frequent charges, we sample two cases for each charge based on the keyword filtering above as the procedural queries. Finally, a total of 100 procedural cases are collected in this section.

3.4 Candidate Pooling

In practice, it is impractical and expensive to annotate the relevance of a query to all other cases. To allocate limit judging resources to more promising cases, depth-k pooling has been applied in many retrieval datasets [2, 5, 23, 24, 26]. This approach involves obtaining a document pool from existing retrieval models and then having annotators label their relevance. For instance, LeCaRDv1 adopts three lexical matching models for pooling and merge the top-100 retrieved cases into the final pool. However, this simple pooling strategy retrieves cases with similar properties, which may does not fully exploit promising cases.

To address this issue, we propose a two-level pooling strategy comprising a retrieval pooling step and a rank pooling step. The retrieval pooling step aims to form the candidate set with diversity by combining sparse lexical matching, dense semantic retrieval, and law article similarity. Subsequently, the ranking pooling further prioritizes the cases in the retrieval pool using runs submitted by previous participants in CAIL2021, with the goal of identifying the most promising cases for annotation.

3.4.1 Retrieval Pooling. For each query, the retrieval pooling step selects 100 candidate cases from the large corpus. It combines sparse lexical matching, dense semantic retrieval, and law article similarity to improve the diversity of candidate sets.

Sparse lexical matching: Sparse lexical matching methods calculate the similarity based on the same words between the query and the candidate. In this section, we employ BM25, a classical lexical matching method, to retrieve the relevant case documents. To be specific, we first apply jieba [4] to split the Chinese sentences into words. Then we remove the stop words and retrieve the top-100 cases for each query from the entire corpus.

Dense semantic retrieval: In recent years, with the development of pre-trained language models, semantic-based dense retrieval models have attracted considerable attention. Dense retrieval models typically employ complex neural networks to encode query and document as h_q and h_d respectively. The semantic relevance scores are then calculated by applying the dot product or cosine similarity to their encoded representations. In general, dense retrieval models can better capture the semantic information of the context through complex interactions. To construct the pool of LeCaRDv2 with dense retrieval models, we apply RoBERTa [20] as the backbone and pre-train it on a legal corpus with whole word mask (WWM) task. The top-100 relevant case documents for each query are retrieved with this model.

Law article similarity: The law article similarity aims to calculate the relevant score by the law article cited in the cases. Since the definition of relevance in the legal field differs from that of the general domain, it is not sufficient to only use text-based similarity methods, i.e., lexical matching and semantic retrieval. Therefore, inspired by inverse document frequency (IDF) [1], we propose Inverse Provision Frequency (IPF) to measure the similarity of cases in terms of the law articles. To be specific, we extract the criminal law articles involved in each case document. Given a specific law article P_i, the IPF value is as follows:

$$IPF_{P_i} = \log \frac{|D|}{freq(P_i, D)} \tag{3}$$

where $|D|$ is the size of the corpus and $freq(P_i, D)$ represents the number of cases containing P_i in corpus D. The design of the IPF is based on the intuition that law articles cited in a large number of cases may not contain information that is important for a particular query. In other words, it cannot provide enough information to determine relevance. The law article similarity of the two cases is calculated by summing the IPF values of all co-cited

[3] These keywords are available in GitHub https://github.com/THUIR/LeCaRDv2.

[4] https://github.com/fxsjy/jieba

law articles. As described above, we use IPF to retrieve the top-100 relevant case documents for each query.

To merge the cases retrieved by the three methods described above into a single pool, we divide them into four groups:

- **Group 1**: top-25 cases retrieved retrieved by any of the above. Note that the total number of cases in this group may be less than 75, since there may be duplicate cases retrieved with different methods.
- **Group 2**: After filtering out the cases in group 1, the remaining cases occurred in the top-100 cases in all three retrieval models.
- **Group 3**: After filtering out the cases in group 1, the remaining cases occurred in the top-100 cases for exactly two retrieval models.
- **Group 4**: After filtering out the cases in group 1, the remaining cases occurred in only one retrieval model.

For each query, the candidate pool contains all cases in group 1. Then, we select cases from group 2 to supplement the candidate pool to 100 cases. If the total number of cases in group 1 and group 2 is less than 100, the leftover cases are sampled from group 3. Similarly, cases from group 4 will be added to the candidate pool if group 3 does not have enough cases. In short, the retrieval pool contains cases with the highest scores in each retrieval model and cases retrieved by multiple models. Given a query, we form a pool of candidate cases with a depth of 100.

3.4.2 Rank Pooling. After retrieval pooling, we conduct rank pooling with the runs provided by participants in the CAIL2021 legal case retrieval track (CAIL2021-LCR). CAIL2021-LCR provides a candidate case pool with a depth of 100 for each query, and participants need to re-rank a limited number of cases per query. These runs are all designed for ranking Chinese criminal law cases, which is consistent with our dataset. To be specific, we collect three award-winning runs in CAIL2021-LCR, which are employed to rank the retrieval case pool. We preserve the top-30 cases of each run and divide them into 4 four groups.

- **Group 1**: top-5 cases ranked with the above three runs. There may be less than 15 cases in this group since some cases may be duplicates.
- **Group 2**: After filtering out the cases in group 1, the remaining cases occurred in the top-30 cases in all three re-rank runs.
- **Group 3**: After filtering out the cases in group 1, the remaining cases occurred in the top-30 cases for exactly two re-rank runs.
- **Group 4**: After filtering out the cases in group 1, the remaining cases occurred in only one re-reank run.

The rank pool contains all the cases in group 1 and replenishes them to 30 in the order of group 2, group 3 and group 4 priority. The rank pool contains cases with the highest scores in each run and ranked in top-30 by multiple runs. Thus we have generated the rank case pool with a depth of 30 for each query, where cases are considered to be more potentially relevant. The rank pool will be available for legal experts to annotate.

3.5 Relevance judgment Criteria

The Supreme People's Court of China has published a guidance document [5] for case relevance under the Chinese legal system. There are three main aspects of relevant cases: basic facts, focus of the disputes, and application of law.

The relevance criteria of LeCaRDv1 focus on basic facts while ignoring the focus of the disputes and application of law. For this dataset, we design a more comprehensive relevance criteria, which follow the official guidance better. Specifically, the advancement of our relevance criteria is reflected in two aspects. First, the annotator needs to judge the Overall Relevance, which follows the official documents strictly. Second, we propose unique relevance evaluation criteria in three aspects: characterization, penalty, and procedure, which comprehensively cover the kinds of information needs of users in judicial practice to conduct legal case retrieval. Next, we describe their definitions in detail.

Characterization Relevance: Characterization Relevance focuses on the basic fact of the case. As with LeCaRDv1, critical factor is employed to measure Characterization Relevance. Critical factor has a substantial impact on the trial of the case. Before annotation, the assessors need to determine whether the query case constitutes a crime and what crime it constitutes. The Characterization Relevance is defined as:

Two cases are defined as relevant in Characterization if the similarity between their critical factors is high.

The critical factors consist of key circumstances and key constitutive elements of the crime (key elements). Key elements are the legal concept abstraction of key circumstances. The different legal elements can lead to different judgments. The Characterization Relevance labeling is in a four-level setting ranging from 1 to 4 with increasing relevance. Detailed descriptions of the relevant labels are as follows:

- **Label-1**: Both key elements and key circumstances are irrelevant.
- **Label-2**: Key circumstances are relevant but key elements are irrelevant.
- **Label-3**: Key elements are relevant but key circumstances are irrelevant.
- **Label-4**: Both key elements and key circumstances are relevant.

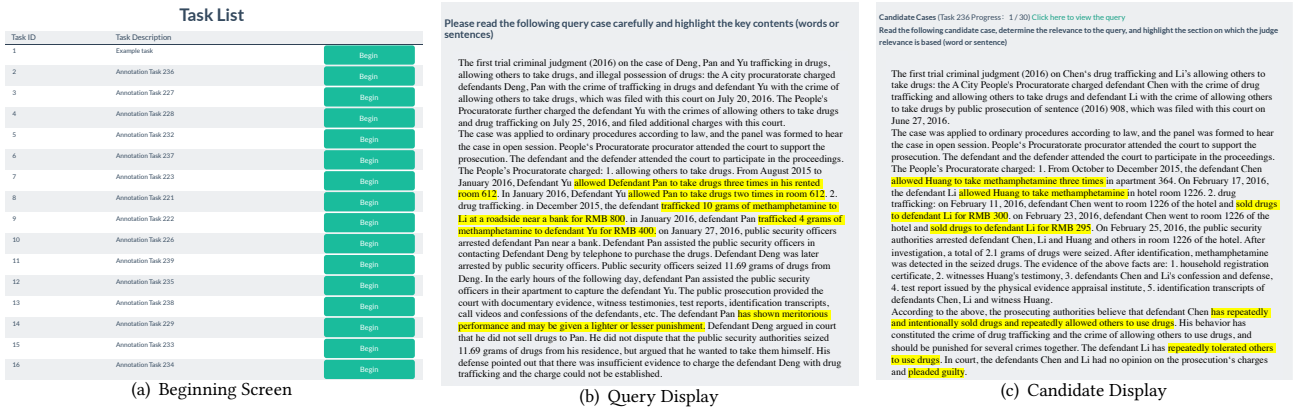
Penalty Relevance: Penalty Relevance focuses on the circumstances of sentencing. In reality, we note that although the basic facts of two cases are similar, the sentences may be different. Before annotation, the annotator needs to determine the subjective factors of the defendant and what the defendant has committed. The Penalty Relevance is defined as:

Two cases are defined as relevant in Penalty if the similarity between their circumstances of sentencing is high.

The circumstances of sentencing consist of the commission of the crime and the individual circumstances of the offender. The commission of the crime includes the criminal pattern, specific means, tools, harmful results and the impact on society of criminal behavior. The individual circumstances of the offender include the offender's age, experience (previous convictions, repeat offenses), attitude after the crime (surrender, covering victims' losses), etc. The Penalty Relevance labeling is also in a four-level setting ranging from 1 to 4 with increasing relevance. Detailed descriptions of the relevant labels are as follows:

- **Label-1**: Both commissions of crime and offender circumstances are irrelevant.

[5] http://www.hncourt.gov.cn/public/detail.php?id=181775



(a) Beginning Screen (b) Query Display (c) Candidate Display

Figure 4: An example of the THUIR search experiment platform. The annotators are required to highlight key content that affects the judgment of relevance.

- **Label-2**: Offender circumstances are relevant but commissions of crime are irrelevant.
- **Label-3**: Commissions of crime are relevant but offender circumstances are irrelevant.
- **Label-4**: Both commissions of crime and offender circumstances are relevant.

Procedure Relevance Procedure Relevance concerns the procedural controversy of the case. Procedural controversy refers to the dispute about the legality of procedural facts in criminal proceedings, which directly affects the application of the law. The Procedure Relevance is defined as:

Two cases are defined as relevant in Procedure if the similarity between their procedural controversy is high.

The procedural controversy includes procedural issues and procedural facts. Procedural issues are disputes involving procedural law, such as the exclusion of illegal evidence, and judicial jurisdiction. Procedural facts are the specific circumstances of the procedural dispute. The Procedure Relevance labeling is in a four-level setting ranging from 1 to 4 with increasing relevance.

- **Label-1**: Procedural issues and procedural facts are irrelevant.
- **Label-2**: Procedural facts are relevant but procedural issues are irrelevant.
- **Label-3**: Procedural issues are relevant but procedural facts are irrelevant.
- **Label-4**: Both procedural issues and procedural facts are relevant.

For Overall Relevance, annotators are required to follow the official guidance criterion and take the characterization, penalty, and procedure relevance as a whole into consideration. We provide two guide cases with relevant score explanations to annotators for better understanding. The Overall Relevance labeling is in a four-level setting which consists of Irrelevant, Somewhat irrelevant, Fairly relevant, and Highly relevant. It is worth noting that there is no explicit mapping function between the Overall Relevance and the sub-relevance. The annotators have discretionary authority and makes judgments based on their knowledge.

3.6 Human Annotation

Our relevance annotators consist of 41 legal experts majoring in criminal law, who have all passed the National Uniform Legal Profession Qualification Examination and are familiar with the cases in our dataset. To ensure the quality of annotation, all annotators first go through several hours of interpretation to ensure a sound understanding of the concept in the criteria. Then, we verify the quality of annotators with several example tasks. The creator of the criteria, who holds a Ph.D. in criminal law, makes corrections according to their annotation in example cases to ensure consistent understanding among all annotators. Only annotators who have passed the training are allowed to perform official annotation. Each annotation task is completed repeatedly by three different annotators. We measure the annotation quality with Kappa [6, 9]. The Kappa value for the Overall Relevance is 0.5190, which indicates that LeCaRDv2 is a high-quality manual labeled dataset. The average of the three annotation results is regarded as the final relevance score. For each annotation task, we pay the legal expert 6.60 dollars.

For efficient annotation, we build an annotation platform to help annotators browse cases and make judgments about their relevance. Figure 4 shows an example of the annotation platform. Figure 4(a) is the beginning screen of the platform, from which the annotator can select the annotation task. Each annotation task contains one query case and thirty candidate cases. The candidate cases are presented in a randomized manner to remove the rank bias of the assessors. In other words, it prevents the annotators from overvaluing the higher-ranked cases and under-valuing the later ones. Figure 4(b) and 4(c) are the query and candidate display screens respectively. During the annotation process, the annotators are also required to highlight key content that affects the judgment of relevance. After careful reading, the annotator needs to make a judgment about four relevance scores between the candidate cases and the query. For Characterization, Penalty, and Procedure relevance, we provide the option with a label=0, meaning that no legal issue is involved between cases in that aspect.

3.7 Data Analysis

In this section, we analyze the different attributes of LeCaRDv2.

Table 1: The statistics of legal class case retrieval datasets.

DATASETS	LeCaRDv1	CAIL2019-SCM	CAIL2022-LCR	COLIEE2020	COLIEE2021	LeCaRDv2
Language	Chinese	Chinese	Chinese	English	English	Chinese
#Queries	107	8264	130	650	900	800
# Candidate cases/query	100	2	100	200	4,415	55,192
Avg.length per case document	8,275	676	2,707	3,232	1,274	4,766
#Avg.relevant case per query	10.33	1	11.53	5.15	4.73	20.89

3.7.1 Data Size. Detailed statistics of LeCaRDv2 and other popular legal case retrieval datasets are shown in Table 1. There are some datasets i.e. LeCaRDv1, CAIL2019-SCM, CAIL2022-LCR, and COLIEE2020 providing a limited number of candidate cases for each query. While COLIEE2021 and LeCaRDv2 require determining relevant documents from the entire corpus, which makes the task more difficult. Compared with LeCaRDv1, LeCaRDv2 is closer to the real-world retrieval task, which requires the model with both efficiency and effectiveness.

From the statistics, we can find that LeCaRDv2 is the largest Chinese legal case retrieval dataset with tens of thousands of annotated data. Benefiting from the candidate case pooling strategy, more promising cases are annotated, which further increases the number of relevant cases.

3.7.2 Data Distribution. Figure 5 shows the distribution of query-candidate pairs with different relevance levels. We can find that most annotated cases are fairly relevant in Overall Relevance, We think it benefits from the large corpus, which can provide sufficient relevant cases, and the candidate pooling strategy, which can determine the more promising relevant cases to annotate. For Characterization Relevance, the number of fairly relevant and highly relevant cases is almost equal. Moreover, it is worth noting that there are some cases that do not involve Penalty Relevance and Procedure Relevance judgment. For Procedure Relevance, since the majority of cases do not involve procedural disputes in China, there are numerous cases with a procedural relevance label of 0. Still, Procedure Relevance is an important component of the relevance in legal cases and deserves further study.

4 EXPERIMENT

In this section, we implement some state-of-the-art models to evaluate LeCaRDv2. We first introduce the benchmark settings and baselines. Then, we report the experimental results and perform a detailed analysis.

4.1 Benchmark Settings

We conduct experiments on state-of-the-art models with zero-shot and fine-tuning settings. Under the zero-shot setting, the model is not trained with any annotated data, which is suitable for legal systems that lack training data. All annotated data in LeCaRDv2 are employed to evaluate. Under the fine-tuning setting, we sampled 20% cases from each charge as the test set. There are 640 queries for training and 160 queries for testing. Since we focus on retrieval performance in large corpus, we adopt recall as the evaluation metric.

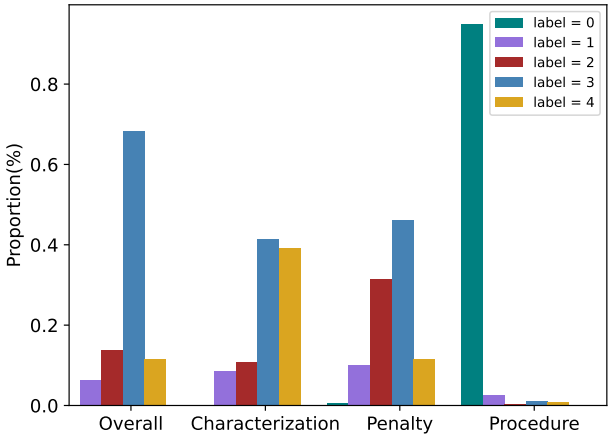


Figure 5: Distribution of cases with different relevance labels. Label=0 means that the query and candidate do not involve this type of relevance.

4.2 Baselines

We adopt three types of widely-used retrieval models as baselines, including Traditional Retrieval Models, Generic Pre-trained Models, and Retrieval-oriented Pre-trained Models. For Pre-trained Models, the dual encoder architecture is applied to retrieve relevant cases from the entire corpus.

- **Traditional Retrieval Models**
  - **BM25** [27] is a robust traditional retrieval model based on word matching.
  - **LMIR** [27] is a highly effective strong baseline mode based on Dirichlet smoothing.
- **Generic Pre-trained Models**
  - **Chinese-BERT-WWM** [7] is a multi-layer transformer trained with Whole Word Mask (WWM) and Next Sentence Prediction (NSP) tasks.
  - **Chinese-RoBERTa-WWM** [7] has the same architecture as Chinese-Bert-WWM, which is trained in enlarged datasets with only WWM task.
  - **BERT_xs** [6] is the Bert specialized in criminal law, which is trained with several million Chinese case documents.
  - **Lawformer** [32] aims to process long legal cases, which employ Longformer [3] as the backbone.
- **Retrieval-oriented Pre-trained Models**
  - **Condenser** [10] designs the skip connection to force information to be integrated into the [CLS] token.

[6] http://zoo.thunlp.org/xs-bert

Table 2: Zero-shot and finetune performance of various baselines on LeCaRDv2. Under zero-shot setting, all queries are applied for evaluation. Under fine-tuning setting, there are 160 queries for testing. The best method in each column is marked in bold.

Models	Zero-shot				Fine-tune			
	R@100	R@200	R@500	R@1000	R@100	R@200	R@500	R@1000
Traditional Retrieval Models								
BM25	**0.6262**	**0.6629**	0.6946	0.7207	**0.6050**	**0.6428**	0.6735	0.7015
QLD	0.5984	0.6576	**0.7065**	**0.7424**	0.5749	0.6354	0.6882	0.7222
Generic Pre-trained Models								
Chinese-BERT-WWM	0.1165	0.1526	0.2184	0.2805	0.3849	0.5026	0.6649	0.7797
Chinese-RoBERTa-WWM	0.3753	0.4739	0.6152	0.7126	0.4136	0.5330	0.6964	0.7998
Bert_xs	0.0453	0.0614	0.0949	0.1343	0.2074	0.2750	0.3935	0.4941
Lawformer	0.2432	0.3040	0.4054	0.4833	0.3651	0.4851	0.6443	0.7629
Retrieval-oriented Pre-trained Models								
Condenser	0.2215	0.2987	0.4321	0.5452	0.3982	0.5003	0.6761	0.7969
coCondenser	0.2255	0.3093	0.4460	0.5514	0.3998	0.5024	0.6861	0.8036
SEED	0.3544	0.4474	0.5745	0.6657	0.4201	0.5437	**0.7160**	0.8132
RetroMAE	0.3193	0.3947	0.5010	0.5821	0.4210	0.5397	0.7093	**0.8174**

- **coCondenser** [11] utilizes unsupervised contrastive learning to warm up the vector space based on Condenser.
- **SEED** [22] applies a weak decoder to enhance the training of the encoder, which achieves state-of-the-art performance in ad-hoc retrieval tasks.
- **RetroMAE** [21] designs a harder decoding process to achieve better retrieval performance.

We use the pyserini toolkit [7] to implement BM25, which with default parameters. For generic pre-trained models, we directly load their checkpoints in huggingface [8]. For retrieval-oriented pre-trained models, we reproduced their work on the legal corpus with their open-source code as there are no available Chinese versions of them. We adopt BERT to initialize retrieval-oriented pre-trained models. Following the previous work [13, 35], negative samples are BM25 negatives and the ratio of positives and negatives is 1:32.

4.3 Experimental Results

The performance of baselines on LeCaRDv2 is shown in Table 2. In the zero-shot setting, we directly measure the performance of all queries with the pre-trained language model. In the fine-tuning setting, we use 640 queries for training and 160 queries for testing. From the experimental results, we can derive the following observation:

- Both in the zero-shot and fine-tuning settings, the traditional retrieval methods show competitive performance on legal case retrieval task.
- Under the Zero-shot setting, generic pre-trained models generally perform worse than traditional retrieval models. Despite training with extensive criminal law data, the performance of BERT_xs is worse than that of BERT. We guess that this is because the pre-training objectives of BERT_xs damage the robustness of the [CLS] embedding, which is not suitable for dense retrieval. Correspondingly, Chinese-RoBERTa-WWM achieves surprising results. This indicates that the next sentence prediction task may not be helpful for dense retrieval.
- Retrieval-oriented pre-training models generally have better performance than generic pre-training models. This indicates that retrieval-oriented pre-training tasks rather than general NLP tasks are more helpful for retrieval.
- With the guidance of labeled data, the performance of pre-trained models is further improved. However, in some metrics i.e. R@100, R@200, BM25 achieves the best performance, which encourages the community to propose more pre-trained language models for legal case retrieval.
- In short, LeCaRDv2 is a challenging retrieval task. Existing pre-trained language models perform worse on legal documents than in the general domain due to length restrictions and different relevance definitions. We are confident that with more test data, LeCaRDv2 can better test the significant effectiveness of the proposed models. It is worth investigating in the future to design better retrieval models in the legal domain.

[7] https://github.com/castorini/pyserini
[8] https://huggingface.co/models

5 CONCLUSION

In this paper, we release LeCaRDv2 as a new and challenging dataset for legal case retrieval. LeCaRDv2 consists of 800 queries covering 50 charges and 55,192 candidate cases, which is one of the largest Chinese legal case retrieval datasets with the widest coverage of criminal charges. We enrich criteria of legal relevance based on LeCaRDv1, which covers characterization, penalty, procedure three aspects and Overall Relevance. Moreover, we propose a novel candidate pooling strategy to identify potential cases with diverse characteristics. We evaluated several competitive baselines on LeCaRDv2. The experimental results show that LeCaRDv2 is a challenging retrieval task and further efforts are needed to promote the development of legal case retrieval. In the future, we will continue to expand the size of this dataset and start focusing on civil law cases. Moreover, we will attempt to extract the highlights of legal experts in determining the relevance of cases for contributing to the community.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (Grant No. 62002194)

REFERENCES

[1] Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* 39, 1 (2003), 45–65.

[2] Piyush Arora, Murhaf Hossari, Alfredo Maldonado, Clare Conran, Gareth JF Jones, Alexander Paulus, Johannes Klostermann, and Christian Dirschl. 2018. Challenges in the development of effective systems for professional legal search. (2018).

[3] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

[4] Paheli Bhattacharya, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2022. Legal case document similarity: You need both network and text. *Information Processing & Management* 59, 6 (2022), 103069.

[5] Charles L Clarke, Nick Craswell, and Ian Soboroff. 2009. *Overview of the trec 2009 web track*. Technical Report. WATERLOO UNIV (ONTARIO).

[6] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[7] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-Training with Whole Word Masking for Chinese BERT. *arXiv preprint arXiv:1906.08101* (2019).

[8] Qian Dong, Yiding Liu, Qingyao Ai, Haitao Li, Shuaiqiang Wang, Yiqun Liu, Dawei Yin, and Shaoping Ma. 2023. I3 Retriever: Incorporating Implicit Interaction in Pre-trained Language Models for Passage Retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 441–451.

[9] Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 33, 3 (1973), 613–619.

[10] Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. *arXiv preprint arXiv:2104.08253* (2021).

[11] Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540* (2021).

[12] Hanjo Hamann. 2019. The German federal courts dataset 1950–2019: from paper archives to linked open data. *Journal of empirical legal studies* 16, 3 (2019), 671–688.

[13] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[14] Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. SAILER: Structure-aware Pre-trained Language Model for Legal Case Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1035–1044. https://doi.org/10.1145/3539618.3591761

[15] Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Zhijing Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2024. BLADE: Enhancing Black-box Large Language Models with Small Domain-Specific Models. arXiv:2403.18365 [cs.CL]

[16] Haitao Li, Qingyao Ai, Xinyan Han, Jia Chen, Qian Dong, Yiqun Liu, Chong Chen, and Qi Tian. 2024. DELTA: Pre-train a Discriminative Encoder for Legal Case Retrieval via Structural Word Alignment. arXiv:2403.18435 [cs.IR]

[17] Haitao Li, Qingyao Ai, Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Zheng Liu, and Zhao Cao. 2023. Constructing Tree-based Index for Efficient and Effective Dense Retrieval. arXiv:2304.11943 [cs.IR]

[18] 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:2305.06812 [cs.IR]

[19] Haitao Li, Changyue Wang, Weihang Su, Yueyue Wu, Qingyao Ai, and Yiqun Liu. 2023. THUIR@ COLIEE 2023: More Parameters and Legal Knowledge for Legal Case Entailment. *arXiv preprint arXiv:2305.06817* (2023).

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[21] Zheng Liu and Yingxia Shao. 2022. RetroMAE: Pre-training Retrieval-oriented Transformers via Masked Auto-Encoder. *arXiv preprint arXiv:2205.12035* (2022).

[22] Shuqi Lu, Di He, Chenyan Xiong, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tieyan Liu, and Arnold Overwijk. 2021. Less is More: Pre-train a Strong Text Encoder for Dense Retrieval Using a Weak Decoder. *arXiv preprint arXiv:2102.09206* (2021).

[23] 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 the 44th international ACM SIGIR conference on research and development in information retrieval*. 2342–2348.

[24] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *choice* 2640 (2016), 660.

[25] Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021. *The Review of Socionetwork Strategies* 16, 1 (2022), 111–133.

[26] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2021. COLIEE 2020: methods for legal document retrieval and entailment. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15–17, 2020, Revised Selected Papers 12*. Springer, 196–210.

[27] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[28] Yunqiu Shao, Haitao Li, Yueyue Wu, Yiqun Liu, Qingyao Ai, Jiaxin Mao, Yixiao Ma, and Shaoping Ma. 2023. An Intent Taxonomy of Legal Case Retrieval. *ACM Trans. Inf. Syst.* 42, 2, Article 62 (dec 2023), 27 pages. https://doi.org/10.1145/3626093

[29] 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 *IJCAI*. 3501–3507.

[30] Olga Shulayeva, Advaith Siddharthan, and Adam Wyner. 2017. Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law* 25, 1 (2017), 107–126.

[31] Marc Van Opijnen and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law* 25 (2017), 65–87.

[32] 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 (2021), 79–84.

[33] 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 the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Taipei, Taiwan,) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2681–2690. https://doi.org/10.1145/3539618.3591874

[34] ChengXiang Zhai. 2008. Statistical language models for information retrieval. *Synthesis lectures on human language technologies* 1, 1 (2008), 1–141.

[35] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. RepBERT: Contextualized text embeddings for first-stage retrieval. *arXiv preprint arXiv:2006.15498* (2020).

[36] Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 3540–3549.