IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 December 2023, accepted 21 January 2024, date of publication 25 January 2024, date of current version 6 February 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3358622*

RESEARCH ARTICLE

# Iterative Self-Supervised Learning for Legal Similar Case Retrieval

**YAO LIU[1,2], TIEN-PING TAN[1], AND XIAOPING ZHAN[3]**
[1]School of Computer Sciences, Universiti Sains Malaysia, Penang 11800, Malaysia
[2]Department of Management and Media, The Engineering and Technology College, Chengdu University of Technology, Leshan 614007, China
[3]School of Law, Sichuan University, Chengdu 610065, China

Corresponding author: Tien-Ping Tan (tienping@usm.my)

This work was supported by the National Social Science of China Fund under Grant 22BFX159.

**ABSTRACT** In the realm of legal artificial intelligence (AI), the spotlight has been cast on its remarkable precision and efficiency, especially in tasks such as similar case retrieval where the identification of pertinent cases in response to a given query is of paramount importance. This task, distinct from traditional text retrieval, presents a set of unique challenges that necessitate the availability of high-quality, annotated datasets to facilitate efficient model training. The intricacies of handling extended queries and candidate documents, coupled with the varied interpretations of similarity, further compound the complexity of this endeavor. This study introduces an innovative training approach, combining dense and sparse retrieval methods. Utilizing a sparse retrieval model, we extract unlabeled data from extensive legal cases. Subsequently, a dense retrieval model screens this data, merging it with labeled data to create pseudo-labeled data, iteratively training until convergence. The results demonstrate exceptional performance in the Chinese law retrieval task dataset, showcasing a notable 3.66% precision enhancement and a substantial 3.62% improvement in mean average precision (MAP). However, the dataset's imbalance across different charges of cases poses a challenge, potentially affecting retrieval performance for long-tailed legal cases. Nonetheless, these outcomes signify accelerated and more efficient retrieval of similar cases for legal professionals. Additionally, they provide high-quality references for non-legal individuals lacking expertise in the field.

**INDEX TERMS** Legal information retrieval, similar case retrieval, iterative training, self-supervised learning.

## I. INTRODUCTION

In recent years, the field of Legal Artificial Intelligence (AI) has raised substantial enthusiasm among legal professionals and technology enthusiasts alike, as it has the potential to revolutionize the legal industry by enhancing the efficiency, accuracy, and accessibility of legal services [1], [2], [3], [4], [5], [6]. This interest has been fueled by the increasing availability of large datasets, and advancements in machine learning algorithms. Additionally, the use of Legal AI has the potential to democratize access to justice, particularly in underserved communities, by providing affordable and efficient legal services. Given these promising developments, it is not surprising that the field of Legal AI is attracting significant attention from legal scholars, practitioners, and policymakers around the world. One of the tasks in Legal AI is legal case retrieval [7], [8], [9], [10]. The legal case retrieval identifies the most relevant or similar cases for a given query, which can be organized into five parts: Procedure, Fact, Reasoning, Decision, and Tail. The formula is expressed as follows: given a query $q$ and a set of candidate documents $D = \{d_i | i = 1, \ldots, n\}$, the class case retrieval task is to extract the most similar $S_{*/q} = \arg\max \{S_{d_i/q} | d_i \in D\}$ documents to $q$ from $d_i$, where $d_i$ is the collection of case documents $d_i \in D$.

In common law jurisdictions such as the United Kingdom and the United States, where courts rely on previous decisions, particularly those of the High Court, as a foundation for current case judgments, the doctrine of the stare decisis [11] is extremely important. whereas not as strained by stare decisis

The associate editor coordinating the review of this manuscript and approving it for publication was Hong-Mei Zhang.

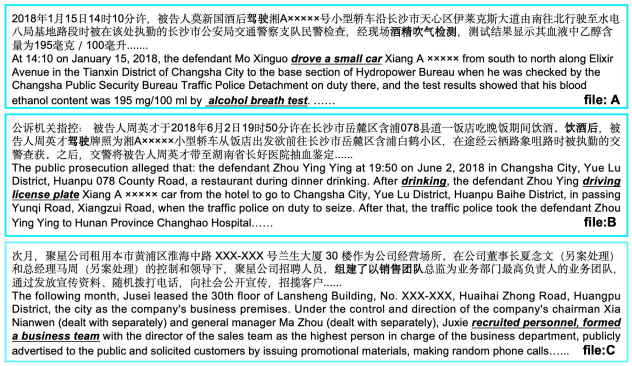



**FIGURE 1. The figure illustrates the three files from the LeCaRD dataset [13], representing actual Chinese legal cases labeled as A, B, and C. File A serves as the foundational case representation for the traffic crime query. In contrast, files B and C represent cases manually annotated by legal experts, denoting similarity and dissimilarity, respectively, to case A.**

in statutory law systems [12] such as China, France, and Germany, knowledge of similar cases remains vital in judicial decision-making and defense strategies. Though precedents are not legally binding, they have a significant impact on legal interpretations and decisions. Searching for comparable cases serves as a guiding tool for the legal community, assisting in the prediction of potential court rulings and the development of effective legal strategies. In a DWI (Driving While Intoxicated) case, for instance (the case presentation in figure1), attorneys can cite jurisprudence from comparable cases to strengthen their argument and anticipate potential court decisions. Lawyers can gauge court tendencies and provide more robust legal advice to clients by referencing similar cases, assisting in understanding legal application and predicting case outcomes.

However, retrieving similar legal cases poses unique challenges distinct from typical information retrieval tasks. Firstly, both query cases and candidate case documents are notably longer in the LeCaRD [13], averaging over 6000 characters each (consult the Table 1). Secondly, defining similarity between legal documents differs from general texts [14]. In the legal domain, identifying similarities demands a deep exploration of the legal facts within the texts. While conventional methods for measuring textual similarity capture semantic similarities, they might miss the intricate nuances specific to the legal domain. Consequently, these methods might falter in identifying critical logical relationships crucial for recognizing similar legal cases. Lastly, the scarcity of well-annotated data amplifies the challenge in training machine learning models. The dataset, as depicted in Table 1, comprises 107 annotated query cases and 10,780 annotated candidate document cases, with a staggering 2 million unannotated cases, representing only a fraction of the total number of current legal cases in China. This issue is not unique to this dataset, as similar challenges are prevalent in datasets across various languages.

To overcome these challenges, we propose a novel training method that combines dense and sparse retrieval methods. We employ a sparse retrieval model to mine unlabeled data from a large-scale legal cases, and then utilize a dense retrieval model for screening, after which it is combined with labeled data to form pseudo-labeled data, and iteratively train the process until the dense retrieval model converges. Our approach significantly improves retrieval efficiency by capitalizing on self-supervised learning. Through the iterative training of hybrid models that integrate sparse and dense retrieval methods, we leverage the computational efficiency of sparse retrieval for handling large-scale data, coupled with the robust semantic comprehension, contextual understanding, and expansiveness of dense retrieval models. In our experiments, this method has showcased remarkable performance in the realm of Chinese legal case retrieval, consistently delivering outstanding results. We summarize the major contributions of the paper as follows:

- We introduce an innovative contrastive learning training approach that simultaneously trains a blend of sparse and dense retrieval models.
- We conduct iterative self-supervised learning using extensive unlabeled data, eliminating the necessity for costly input from legal experts to label the data. This approach yields superior performance in similar case retrieval
- We validated the proposed method on two similar case retrieval datasets, both of which yielded significant performance improvements.

The subsequent sections of this article will comprehensively explore various facets. The Related Works (II) section will delve into existing literature and studies pertinent to information retrieval models. Following this, our proposed Iterative Self-Supervised Training for Legal Similar Case Retrieval (III) section will meticulously present our innovative framework in detail. Subsequently, the Experiments (IV) section will intricately detail the experimental setup, including tools, datasets, parameters, and procedures utilized to validate our proposed method. This will be followed by the Results and Discussion (V) section, where we'll comprehensively present and analyze the experiment outcomes. Finally, the Conclusion (VI) section will synthesize the findings and discuss the broader implications of our proposed approach.

## II. RELATED WORK

Over the past several decades, a wide range of information retrieval models were developed for ad-hoc information retrieval. These models can be broadly divided into two categories: sparse retrieval models and dense retrieval models.

### A. SPARSE RETRIEVAL MODELS

Sparse retrieval models represent a pivotal class of algorithms within the domain of information retrieval (IR), prioritizing the discernment of pertinent documents through meticulous analysis of specific keywords or terms' frequency and significance. Diverging from dense models, which encapsulate

**TABLE 1. Statistic of labeled and unlabeled data.**

| DATASET | LeCaRD | COLIEE2020 | COMMON LAW CASE | CIVIL LAW CASE |
|---|---|---|---|---|
| *Language* | chinese | english | english | chinese |
| *# Query case* | 107 | 130 | - | - |
| *# Candidate case* | 10,780 | 26,000 | 50,000 | 2,000,000 |
| *Avg. length per query case* | 445 | 613 | - | - |
| *Avg. length per candidate case* | 6,319 | 3,232 | 3,015 | 7,215 |

documents and queries as continuous vectors, sparse models hinge on discrete representations. These representations typically manifest as vectors predominantly populated by zero values, with non-zero elements delineating the presence and weightage of distinct keywords. This inherent sparsity underpins efficient retrieval by accentuating the most influential keywords, thereby rendering sparse models exceptionally apt for managing expansive datasets, such as those found in legal case documents and scholarly search [15]. The Bag of Words (BoW) models predominantly operate at the term level, wherein a document is portrayed as an amalgamation of its constituent words devoid of any consideration for word order or grammatical structure. Eminent algorithms based on BoW include the Term Frequency-Inverse Document Frequency (TF-IDF) [16], leveraging the raw frequency of a term within a document (TF) and its inverse document frequency (IDF) to compute scores for each term-document pair. The Vector Space Model for Information Retrieval (VSMIR) [17] portrays documents and queries as vectors within a high-dimensional space, each dimension corresponding to a term. The Best Matching 25 (BM25) [18] assigns a score to each document based on factors including term frequency in the document and query, inverse document frequency (IDF) of the term, and document length. Lastly, the Language Model for Information Retrieval (LMIR) [19] utilizes statistical language models to estimate the probability of generating a query from a document, wherein higher probabilities signify greater relevance.

### B. DENSE RETRIEVAL MODELS

On the other hand, dense retrieval models use a dual encoder architecture to learn document and query embeddings. Unlike sparse retrieval models, dense retrieval models capture word order and contextual information by considering the entire word sequence in the document or query. Based on our reading, we can divide dense retrieval models into two categories: single-vector representation, where the entire input text is represented by a single vector, and multi-vector representation, where the input query and candidate documents can be represented by multiple contextual vectors. In the single-vector representation category, Clinchant [20], Gillick [21] used pre-trained word vectors to represent the unique representations of queries and candidate documents, achieving better practical results than symbolic-based retrieval models. More recently, pre-trained models with better representation capabilities, such as DPR [22] and RepBERT [23], have been widely used as the encoders of dense retrieval models in the past three years. However, this simple structure may cause serious information loss during the encoding of documents since the queries are agnostic. In the multi-vector representation category, ColBERT [24] and Gao [25] used the MaxSim operator to compute the similarity between the query and candidate documents after encoding, while Luan [26] designed a method to mimic the queries on each of the documents by an iterative clustering process and represent the documents by multiple cluster centroids queries. Dense retrieval models exhibit robust semantic comprehension, resilience in contextual understanding, and scalability. However, they are hampered by high training costs and vulnerability to poor learning in zero-shot and few-shot scenarios. Acquiring an excessive amount of labeled domian text data poses a significant challenge, especially within the legal domain. Given these constraints, we advocate for the adoption of a hybrid sparse and dense retrieval model tailored specifically for legal text-based case retrieval tasks.

### C. LEGAL SIMILAR CASE RETRIEVAL

Legal information retrieval can also be divided into two categories: sparse retrieval models and dense retrieval models. In the COLIEE 2019 competition, Wehnert [27] integrated BM25 scores with word centroid distances from word embeddings. This fusion, followed by applying a similarity threshold to varying document retrieval numbers per query, resulted in a refined set of analogous cases. Leveraging word embeddings and textual entailment, this approach adeptly resolves keyword mismatches, captures contextual nuances, manages complex queries, and efficiently operates with limited labeled data. Bin [28] introduced a recommendation model tailored for prevalent legal text scenarios, employing a thematic model approach. Their methodology aimed to refine the word probability distribution beneficial for legal text representation while mitigating the influence of frequently occurring yet irrelevant words. They achieved this through the utilization of regular and weighted TF-IDF from analogous texts. This strategy, leveraging latent topics instead of keywords, enables the system to overcome keyword mismatches, capture contextual intricacies, handle complex queries, and efficiently operate even in scenarios with limited labeled data. Wang [29] presented a legal text similarity

measure for Chinese legal judgments based on a topic modeling approach. The method utilizes TF-IDF, LDA, and Labeled Latent Dirichlet Allocation (LLDA) processing, and has been shown to be effective in handling long texts. Nonetheless, these studies do not consider contextual information. As a result, some later studies have employed a word2vec vector representation to compute case similarity. According to research conducted by Deng [30], utilizing a combination of word2vec, doc2vec, and TF-IDF algorithms to compute case similarity can yield improved outcomes for class case retrieval. Although these methods have yielded favorable results, the traditional sparse method that relies on semantic information of lexical items and TF-IDF has limitations, such as incomplete information of keyword vectors and lack of syntactic information. To overcome these shortcomings, Li [31] proposed an improved method of calculating keyword vectors using bipartite graphs and incorporating syntactic information to calculate document similarity. Additionally, a dual network computation model with an attention mechanism was designed.

Recent research has shown that the use of dense information representation, such as BERT [32], [33], has gained significant attention in the domain of scholar search and legal case retrieval. The BERT-PLI [34], a novel model that utilizes BERT to capture semantic relationships at the paragraph level and then uses these interactions to infer the relevance between two cases. To adapt the BERT model to legal scenarios, they fine-tuned it on a small-scale case law dataset and used a paragraph-level framework to reduce computational costs. BERT-LF [35], proposes a similarity case retrieval method based on legal facts. The model combines topic and legal entity facts to enhance the document representation vector's suitability for legal scenarios. Moreover, it uses a BERT-based paragraph aggregation method to encode contextual semantic information and address the issue of long texts. Fang [36] proposes three different data enhancement methods, namely truncation, double-dropout, and prompting, to improve similar case matching results. They combine BERT [37] and TextGCN [38] to achieve more effective comparative learning results in a simple and efficient manner.These models excel at analyzing word relationships and contextual nuances within legal texts, retrieving cases that are specifically relevant to specific legal aspects. They have a thorough understanding of legal concepts, allowing for the accurate retrieval of related cases. Additionally, their proficiency in modeling sequential patterns enhances comprehension of legal arguments, thereby refining case retrieval accuracy in intricate tasks. Although dense retrieval models offer robust semantic understanding, contextual depth, and scalability, they suffer from drawbacks like high training expenses and limited learning capability with scant data'a common challenge in the legal sphere where obtaining abundant labeled legal text data is difficult. Given these considerations, we strongly advocate for a hybrid retrieval model that merges dense and sparse retrieval methods. This approach capitalizes on the strengths of both paradigms, navigating the complexities of legal text-based case retrieval efficiently.

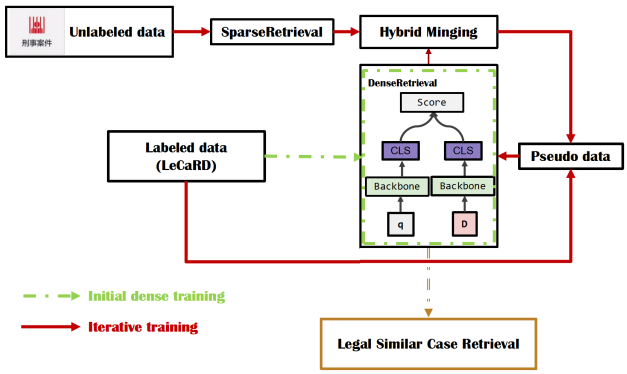



**FIGURE 2.** Overview of iterative self-supervised training for legal similar case retrieval. The green dashed line symbolizes the process of training the backbone model with labeled data, while the red line indicates iterative training.

Despite significant advances regarding employing information retrieval techniques in the legal domain utilizing typical and deep learning approaches, the primary focus has been on exploiting a small amount of labeled data. Regrettably, this technique ignores the potential benefits of exploiting large amounts of unlabeled data and fails to capitalize on the inherent advantages of both sparse and dense retrieval approaches. address this gap, our strategy maximizes the utilization of a small yet high-quality set of labeled data alongside an extensive reservoir of unlabeled legal text data. We present a hybrid sparse and dense retrieval model that uses a contrasting learning self-supervised iterative strategy to enhance performance and leverage the characteristics of both retrieval paradigms.

## III. PROPOSED ITERATIVE SELF-SUPERVISED TRAINING FOR LEGAL SIMILAR CASE RETRIEVAL

Self-supervised learning is a machine learning approach that ingeniously employs implicit signals or structures present in the input data as labels instead of relying on human-annotated data. This technique drastically removes the labeling costs and enables the handling of large datasets where human labeling is unfeasible. Typically, self-supervised learning takes advantage of the structure of the input data as the label for training the model. For example, in language modeling, a model can perform self-supervised learning by predicting missing words from an extensive corpus of text data. Furthermore, models trained through self-supervised learning can also be used for transfer learning, allowing them to be fine-tuned for other tasks with great efficacy.

In this study, we present an iterative hybrid pseudo-labeled retrieval model designed to enhance the retrieval performance of Chinese legal cases, and the overview is shown in figure 2. We leverage a combination of dense and sparse retrieval

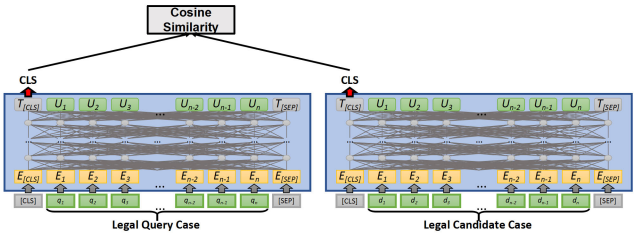



**FIGURE 3.** The query case and the candidate case are individually fed into a Dual Tower pre-training model for encoding. Subsequently, the [cls] embeddings derived from both encoding processes undergo a cosine similarity operation to yield their respective similarity scores.

models to retrieve similar legal cases. The proposed approach involves a three-step process, starting with the training of a bi-encoder dense retrieval model on a labeled Chinese legal case dataset, LeCaRD [13], to establish an initial dense retrieval model, indicated by the GREEN DASHED line in figure 2. Subsequently, a sparse retrieval model is trained on a large unlabeled Chinese legal case dataset, and the top-j (j is the number of candidates with the highest evaluations from unlabeled cases. e.g., j=100) candidate cases are retrieved for each query. Next, we employ the initial dense retrieval model from the first step to retrain against the candidate cases obtained from the sparse retrieval results in the second step. This refined process allows us to extract the top-k (k is the number of highest scoring documents extracted from candidate instances by unsupervised training. e.g., k=30) candidate cases exhibiting the highest similarity scores. The selected top-k cases are then merged with the labeled data, forming hybrid pseudo-labeled case data. The iterative training phase commences by repeatedly updating the initial dense retrieval model. In each iteration, the model recalls the top-k cases with the highest similarity scores from the top-j cases in the unlabeled dataset, indicated by the RED line in figure 2. These retrieved cases are integrated into the pseudo-labeled case data alongside the labeled candidate cases. This iterative training process continues until convergence is achieved, ensuring the dense retrieval model progressively improves and adapts to the unique characteristics of the legal case domain.

### A. DENSE PASSAGE RETRIEVAL MODEL WITH CONTRASTIVE LEARNING

We apply a dense retrieval model to legal similar case retrieval. The dense retrieval model consists of dual encoder to encode query case documents and their corresponding candidate case documents. Specifically, the encoder consists of a pre-trained model, which is used to obtain word embeddings for the query and candidate cases. and subsequently measuring their similarity, shown in figure 3.

As widely acknowledged, dense contrastive learning seeks to generate meaningful representations by attracting semantically similar instances while repelling dissimilar ones. Our study operates on a dataset $D(q, d) = \{q, d^+, d^-\}$, where $d^+$ represents the similar case corresponding to query q in the labeled dataset, and $d^-$ signifies a case that is dissimilar to the query. Inspired by the comparison framework proposed by [22], we employ a objective with mini-batch, defining the objective function as follows:

$$\ell_{(q,d_q^+)} = -\log \frac{e^{sim(h_q,h_q^+)/\tau}}{\sum e^{sim(h_q,h_q^-)/\tau}} \tag{1}$$

Here, $q$, $d^+$, and $d^-$ denote the representations $h_q$, $h_q^+$, and $h_q^-$, respectively, while $\tau$ serves as a temperature hyperparameter improving the performance of the model by regulating the diversity while ensuring accuracy. To quantify similarity, we adopt the cosine similarity, defined by Equation 2:

$$sim(h_1, h_2) = \frac{h_1^T h_2}{\|h_1\| \cdot \|h_2\|} \tag{2}$$

Our approach utilizes the RoBERTa [39] and Longformer [40] model as the backbone encoder to encode all cases within the dataset, followed by fine-tuning of all parameters using the contrast learning objective equation 1.

### B. UNSUPERVISED HYBRIDS DATASET MINING

In this study, we propose an unsupervised hybrid data mining approach to enhance the retrieval performance of legal case documents. The methodology involves fine-tuning the training query document and its corresponding candidate document from the labeled dataset, LeCaRD [13], using a dense paragraph retrieval training method with a pre-training model, thus obtaining an initial dense retrieval model.

Specifically, the training query document $q$ and its corresponding candidate document $d$ from the labeled dataset (LeCaRD) are encoded using the pre-training model. The similarity between the query document and the candidate document is computed through the dot product of their encoding results, as represented by:

$$sim_{Dense}(q, d) = \mathbf{q} \cdot \mathbf{d}^T \tag{3}$$

Here, $\mathbf{q}$ represents the embedding vector representation of the query document after encoding with the pre-training model, and $\mathbf{d}$ signifies the embedding vector representation of the candidate set of documents corresponding to the query document after encoding with the training model.

To address the challenge of expensive legal manual labeling, we employ a combination of SPARSE and DENSE mining techniques to obtain pseudo-labeled data. Initially, we train the dense paragraph retrieval model using open-source annotated data and acquire dense retrieval scores for the corresponding queries. Subsequently, we leverage the BM25 algorithm to extract a candidate case training set for each query document from a large amount of unlabeled data.

$$sim_{Sparse}(q, d) = \sum_{i}^{n} IDF(q_i^*) \cdot \frac{f_i \cdot (k_1 + 1)}{f_i + k_1 \cdot (1 - b + b\frac{dl}{avgdl})} \tag{4}$$

The $IDF(q_i)$ formula is denoted as:

$$IDF(q_i) = \log \frac{N - n(Q_i) + 0.5}{n(Q_i) + 0.5} \quad (5)$$

Here, $N$ denotes the number of documents to be retrieved, and $n(Q_i)$ is the number of documents containing $Q_i$. Moderating factors $k_1$ and $b$ are set to 2 and 0.75, respectively, in this study. Additionally, $f_i$denotes the frequency of $Q_i$ appearing in candidate documents, $dl$ represents the length of candidate documents, and $avgdl$ is the average length of all documents. Following Equation (6),

$$S_{(q,d)} = \lambda \left\| sim_{Dense}(q, d) \right\| + (1 - \lambda) \left\| sim_{Sparse}(q, d) \right\| \quad (6)$$

we obtain the mixed score of candidate documents with unlabeled data for each query. Based on this score, we identify high-quality positive samples and combine them with the original labeled data to form the pseudo-labeled candidate data. The similarity score is denoted by $\|sim(\bullet)\|$, and $\lambda$ represents the smoothing balance index.

**Algorithm 1** Unsupervised Hybrid Data Mining

**Input: (1) labeled legal similar case retrieval dataset $D_L$, (2) unlabeled legal case dataset $D_W$, a pre-trained model $\Theta$**

function Mine($\Theta, D_L, D_W$)
  Initial model $\Psi \leftarrow Training(\Theta, D_L)$ ▷ Train an initial dense retrieval model with a backbone model on data labeled by legal experts.
  for every query do
    $S_1 \leftarrow Sparse(q, D_W)$ ▷ Equation 4 Construct sparse retrieval similarity scores for query cases based on unlabeled data $D_W$, and then extract the top 100 sparse candidate documents $D_W^*$ for each query.
    $S_2 \leftarrow Dense(\Psi, D_W^*, q)$ ▷ Equation 3 Obtain dense retrieval similarity score from sparse retrieval candidate data $D_W^*$ and use faiss [41] for dense indexing of documents.
    $S = Score(S_1, S_2)$ ▷ Equation 6 Integrate the results of both sparse and dense retrieval models.
    $\widehat{D_W} = Sort(S, D_W^*)$ ▷ Based on the score of the $D_W^*$ data for each query, the top 30 candidate fusion documents $\widehat{D_W}$ are ranked.
    $D_L^* = D_L + \widehat{D_W}$ ▷ Pseudo-labeled datasets $D_L^*$ are formed by mixing labeled data candidate documents $D_L$ and unlabeled fusion data candidate documents $\widehat{D_W}$ for each query.
  end for
  return $D_L^*$
end function

### C. ITERATIVE MINING AND RETRIEVAL TRAINING

Building on the unsupervised mining capability of the pretrained model (Algorithm 1), we propose an iterative training framework (Algorithm 2) to improve the pretrained models for both mining and downstream legal document retrieval task. The iterative process involves training the initialized dense retrieval model on a sparse set of retrieval case candidates and mining relevant positive samples corresponding to each query. Additionally, the labeled sample data is incorporated to form pseudo-labeled data. This iterative training process is repeated multiple times until the retrieval model converges. Notably, the initial retrieval model utilized in hybrid mining is the same dense retrieval model that undergoes iterative training, as depicted in Algorithm 1 and Algorithm 2.

**Algorithm 2** Iterative Self-Supervised Learning for Legal Similar Case Retrieval

**Input: (1) labeled legal similar case retrieval dataset $D_L$, (2) unlabeled legal case dataset $D_W$, (3) a pre-trained model $\Psi$, (4) the totle number of iterations $T$**

Initial $\Theta \leftarrow \Psi, t = 0$ ▷Initialize the backbone model.
while $t < T$ do
  $D_L^* \leftarrow Mine(\Theta, D_W, D_L)$ ▷ Algorithm 1
  $\Theta \leftarrow BackboneRetrieval(\Psi, D_L^*)$ ▷ Iteratively train and update the backbone retrieval model $\Theta$.
end while

## IV. EXPERIMENTS

### A. DATASET

We conduct our experiments on two legal case retrieval benchmarks and two extra datasets as external unlabeled data. The statistics are shown in Table 1.

- **LeCaRD** [13] is the first Chinese legal case retrieval open-source dataset, by Tsinghua University, which serves as annotated data for training and testing.
- **CHINESE UNLABELED DATASET** is another dataset obtained from the 'China Judgment Online'.[1]
- **COLIEE2020** [42] is the official dataset provided by COLIEE2020.[2] Each query has 200 candidates. The dataset provides several gold labels for each query.
- **ENGLISH UNLABELED DATASET** is the dataset used to validate the English dataset for this study's approach. We collected our extensive corpus of case documents from the U.S. federal and state courts.[3]

The LeCaRD [13] dataset comprises 107 query cases and 10,700 candidate cases meticulously curated from over 43,000 Chinese criminal judgments. For our experimentation, 20 query documents along with their corresponding candidate documents were chosen as test data, while the remaining 87 query documents and their corresponding candidates were allocated for training. Additionally, we utilized a dataset obtained from the 'China Judgment Web.' These unlabeled

[1]https://wenshu.court.gov.cn/
[2]https://sites.ualberta.ca/ rabelo/COLIEE2020/
[3]https://case.law/

cases served in an iterative self-supervised learning process, encompassing both subjective and objective assessments to represent a diverse array of scenarios. This dataset, containing over 2,000,000 publicly available criminal legal judgments spanning 2020 to 2022, underwent thorough cleaning and categorization based on specific penal charges, detailed in Table 2. This rigorous process ensured inclusivity across typical and contentious cases, bolstering the robustness of our model. Moreover, we incorporated the dataset officially provided by COLIEE2020 to validate our proposed methodology. As an external dataset, we collected 500,000 cases from the U.S. federal and state courts, augmenting our experimental scope.

### B. BASELINE METHODS AND EXPERIMENTAL SETTINGS

We compared our model experimentally with the following baseline models: conventional sparse-based retrieval models and neural network-based models. For the conventional sparse-based retrieval models, we used three popular models, TF-IDF [43], BM25 [18], and LMIR [19], with all parameters set to their default values. For the pretrained encoder, we employ BERT [37], RoBERTa [39] and BERT-LF [35], which integrated topic and legal entity facts to enhance the document representation vector's suitability for legal scenarios, utilized a BERT-based paragraph aggregation technique to encode contextual semantic information and overcome the limitation of long texts. We also compared our results to a baseline system, Lawformer [44]. It is a binary classifier, where it determines if a candidate case is relevant to the query case. The fine-tuning batch size was set to 32 and the learning rate to $10^{-5}$ for all models. The maximum length of the query and candidate was set to 509 and 3,072, respectively, and all tokens in the query case were used for the global attention mechanism. We utilized the PyTorch framework to train our models, with the RoBERTa [39] and Longformer [40] model serving as the backbone feature encoder. We employed 2 Tesla V100 GPUs for fine-tuning the model, with a warmup ratio of 0.1, a learning rate of $2e^{-5}$, a weight decay of 0.01, a batch size of 16. For comparison purposes, we also conduct experiments using the same GPU environment as reported in previous works.

## V. RESULTS AND DISCUSSION

In evaluating the performance of our models, we adopt metrics that are widely used in the literature, which include precision and ranking scores. The precision metrics include precision at rank 5 (P@5), precision at rank 10 (P@10), and mean average precision (MAP), which measure the fraction of relevant documents in the top k returned by the model. The ranking metrics, on the other hand, include normalized discounted cumulative gain (NDCG) [45], in Table3, we denote as N@5, N@10, N@20, N@30, which consider both the relevance of the retrieved documents and their rank in the returned list. For COLIEE task, we report the mean reciprocal rank (MRR), Precision@5, Recall@5, and F1 score.

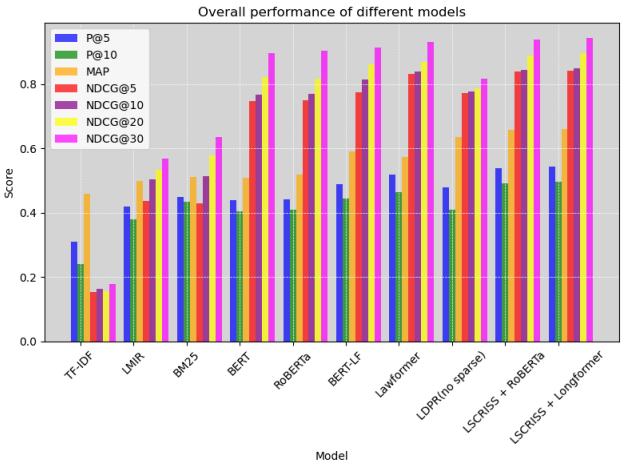



**FIGURE 4. Comparison of the performance of different retrieval models on the test dataset.**

### A. OVERALL RESULTS

As shown in the Figures 4 and Table 3, our model achieves significant performance improvements over other baseline models on both Chinese legal case retrieval datasets. We argue that our model can utilize more unlabeled data than other baseline models. We use the efficient classic sparse model BM25 to augment positive samples and iteratively train a pre-trained model based on a dual-tower architecture. In addition, since case retrieval models require the model to calculate the similarity between the input query case and the candidate cases, traditional models such as BM25 [18] and LMIR [19] cannot well understand the semantic information of text. Some pre-trained baseline models, BERT [37] and BERT-LF [35], cannot well read the entire text due to the limitation of input tokens, resulting in only calculating the similarity between two cases in a certain section or summary, which leads to their poor performance on the test set. It's noteworthy that our training method produces impressive results on two distinct pre-trained models. Moreover, the method tailored for longer texts attains a state-of-the-art performance level.

### B. MODEL ABLATION

In our ablation experiments, we utilize the LeCaRD [13] dataset to conduct a comprehensive evaluation of our proposed hybrid retrieval model. These experiments are specifically designed to analyze and discern the influence of several pivotal factors on the performance of this retrieval model. We're specifically investigating the effects of integrating hybrid sparse retrieval models, which mine data from unlabeled datasets, into the iterative training of pre-trained models. To illustrate, we're comparing two scenarios: *LDPR (no sparse)* and *LDPR + BM25*, which represent instances with and without the inclusion of these hybrid sparse retrieval models, respectively. Moreover, we're examining the influence of different sparse retrieval models, such as *LDPR + TF-IDF*, on the overall enhancement of the

**TABLE 2. Statistic of different charges for unlabeled data.**

| Charge name | Number | Charge name | Number |
|---|---|---|---|
| *Larceny* | 534,324 | *Intentionally destroying possessions* | 16,703 |
| *Dangerous driving crime* | 418,757 | *Bribery* | 14,297 |
| *Intentional injury crime* | 263,809 | *Illegal business crime* | 14,082 |
| *Traffic accident crime* | 195,926 | *Contract fraud crime* | 13,449 |
| *Fraud* | 81,324 | *Extortion crime* | 12,422 |
| *Providing venues for drug users* | 73,720 | *Affray crime* | 12,177 |
| *Defiance and affray crime* | 70,962 | *Corruption crime* | 11,222 |
| *Robbery* | 40,087 | *Negligence causing death crime* | 7,489 |
| *Casino crime* | 39,503 | *Rape crime* | 6,990 |
| *Disrupting public service crime* | 29,643 | *Offering bribes crime* | 5,915 |
| *Credit card fraud crime* | 25,803 | *Crime of refusing to execute judgments or orders* | 3,739 |
| *Illegal detention crime* | 23,802 | *Negligently causing serious accident crime* | 3,700 |
| *Deforestation crime* | 22,948 | *Crime of unlawful intrusion into residence* | 2,449 |
| *Gambling crime* | 16,931 | *Nongovernmental staff bribery crime* | 2,396 |

**TABLE 3. Overall results.**

| Model | LeCaRD | | | | | | | COLIEE2020 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P@5 | P@10 | MAP | N@5 | N@10 | N@20 | N@30 | Precision | Recall | F1_score | MRR@10 | MRR@50 |
| *TF-IDF* | 0.310 | 0.24 | 0.459 | 0.154 | 0.164 | 0.157 | 0.178 | 0.468 | 0.568 | 0.507 | 0.719 | 0.722 |
| *LMIR* | 0.420 | 0.38 | 0.498 | 0.437 | 0.503 | 0.534 | 0.568 | 0.412 | 0.552 | 0.499 | 0.713 | 0.731 |
| *BM25* | 0.450 | 0.435 | 0.512 | 0.430 | 0.513 | 0.578 | 0.636 | 0.473 | 0.571 | 0.518 | 0.787 | 0.790 |
| *BERT* | 0.440 | 0.405 | 0.510 | 0.747 | 0.767 | 0.821 | 0.897 | 0.452 | 0.556 | 0.791 | 0.791 | 0.793 |
| *RoBERTa* | 0.443 | 0.410 | 0.518 | 0.751 | 0.769 | 0.818 | 0.903 | 0.463 | 0.583 | 0.762 | 0.760 | 0.762 |
| *BERT-LF* | 0.490 | 0.445 | 0.592 | 0.775 | 0.815 | 0.863 | 0.915 | - | - | - | - | - |
| *Lawformer* | 0.519 | 0.464 | 0.573 | 0.831 | 0.840 | 0.870 | 0.932 | - | - | - | - | - |
| *LDRP (no sparse)* | 0.480 | 0.41 | 0.635 | 0.773 | 0.777 | 0.787 | 0.818 | 0.531 | 0.703 | 0.610 | 0.876 | 0.879 |
| *LSCRISS + RoBERTa* | 0.538 | 0.491 | 0.658 | 0.839 | 0.845 | 0.890 | 0.939 | 0.543 | 0.711 | 0.613 | 0.881 | 0.883 |
| *LSCRISS + Longformer* | 0.543 | 0.497 | 0.661 | 0.842 | 0.849 | 0.896 | 0.943 | 0.546 | 0.716 | 0.617 | 0.887 | 0.889 |

retrieval model's performance. This involves an exploration of how different sparse retrieval methods contribute to the effectiveness of the broader retrieval model. Simultaneously, we're comparing the performance outcomes of our proposed method across two distinct backbone models, known as *LSCRISS + RoBERTa* and *LSCRISS + Longformer*. These comparisons are detailed in Table 4 and visually represented in Figure 5. Overall, our findings showcase the effectiveness of our iterative approach in significantly improving the performance of dense paragraph retrieval models. This research is pivotal in advancing the accuracy and efficiency of information retrieval systems, particularly within the domain of legal text-based case retrieval.

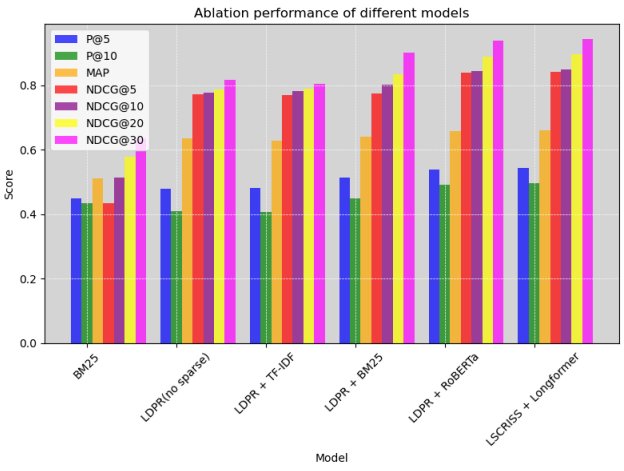



**FIGURE 5. Ablation experiment results on the test dataset.**

### C. HYPERPARAMETER ANALYSIS

Throughout our experiments, we maintained a fixed number of iterations, precisely 5 in total. Convergence, a pivotal aspect of our investigation, was assessed through the performance metrics P@5, MAP, NDCG@5, and NDCG@10, meticulously computed upon the final model. The result delineates in Table 5 and Figure 6. The zenith of model

**TABLE 4.** Ablation results.

| Model | P@5 | P@10 | MAP | NDCG@5 | NDCG@10 | NDCG@20 | NDCG@30 |
|---|---|---|---|---|---|---|---|
| *BM25* | 0.450 | 0.435 | 0.512 | 0.435 | 0.513 | 0.578 | 0.636 |
| *LDPR( no sparse)* | 0.480 | 0.41 | 0.635 | 0.773 | 0.777 | 0.787 | 0.818 |
| *LDPR + TF-IDF* | 0.482 | 0.408 | 0.628 | 0.769 | 0.783 | 0.789 | 0.805 |
| *LDPR + BM25* | 0.513 | 0.45 | 0.641 | 0.774 | 0.802 | 0.834 | 0.902 |
| *LSCRISS + RoBERTa* | 0.538 | 0.491 | 0.658 | 0.839 | 0.845 | 0.89 | 0.939 |
| *LSCRISS + Longformer* | 0.543 | 0.497 | 0.661 | 0.842 | 0.849 | 0.896 | 0.943 |

**TABLE 5.** Iterative times results.

| Model | P@5 | MAP | NDCG@5 | NDCG@10 |
|---|---|---|---|---|
| *LSCRISS Iter 1* | 0.513 | 0.641 | 0.774 | 0.802 |
| *LSCRISS Iter 2* | 0.524 | 0.649 | 0.781 | 0.831 |
| *LSCRISS Iter 3* | **0.543** | **0.661** | **0.842** | **0.849** |
| *LSCRISS Iter 4* | 0.531 | 0.647 | 0.828 | 0.837 |
| *LSCRISS Iter 5* | 0.527 | 0.621 | 0.820 | 0.829 |

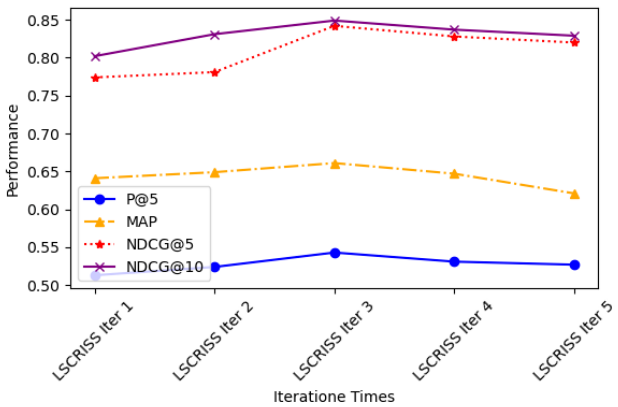



**FIGURE 6.** Experiment results of iterative times on the test dataset.

**TABLE 6.** Pseudo-data results.

| Model | P@5 | MAP | NDCG@5 | NDCG@10 |
|---|---|---|---|---|
| *w-top200* | 0.536 | 0.648 | 0.834 | 0.838 |
| *w-top100* | **0.543** | **0.661** | **0.842** | **0.849** |
| *w-top50* | 0.531 | 0.642 | 0.83 | 0.836 |
| *w-top30* | 0.517 | 0.629 | 0.828 | 0.831 |

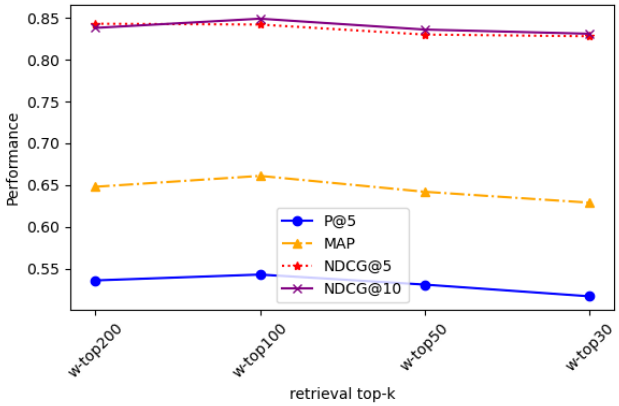



**FIGURE 7.** Experimental results of pseudo-data on the test dataset.

performance unequivocally manifests itself after the third iteration, marking the epoch of peak achievement. Subsequent iterations fail to yield appreciable enhancements, unequivocally signaling the attainment of convergence. The genesis of this noteworthy convergence phenomenon can be ascribed to our choice of employing the pre-trained model as the foundational backbone of our study. An integral facet of our research revolves around the adept integration of pseudo-labeled data, obtained through self-supervised learning and mining. This strategic integration expedites the convergence process, with its most pronounced effects conspicuously emerging after the third iteration.

Furthermore, in the composition of our pseudo-labeled data, we meticulously extracted top-k samples (where k=200, 100, 50, 30) from the unlabeled dataset, subsequently amalgamating them with the labeled data. Our discerning experimental findings, shown in Table 6 and Figure 7, manifestly corroborate that the selection of the top-100 samples engenders the zenith in both model performance and efficacy.

A rigorous sequence of ablation experiments with diverse hyperparameter settings firmly substantiates our assertion that the model achieves its pinnacle when trained with the top-100 unlabeled data samples. It is imperative to underscore that prior to the selection of unlabeled data, we judiciously deployed an efficient BM25 model for recall and screening. Furthermore, we duly acknowledge the impact of data imbalance within the unlabeled dataset, which begets disparities in the retrieval of case data by the sparse retrieval model, thereby substantiating the consequential influence of varying unlabeled data selection criteria on model performance.

## VI. CONCLUSION

In this study, we propose a data-augmented pseudo-labeled iterative dense passage retrieval model for the task of similar legal case retrieval. Our model was initially trained using the

annotated LeCaRD dataset and then underwent an iterative training process in which a mixture of BM25 and the initial model was used until convergence was achieved. This experiment leverages publicly available data from the 'China Judgment Online', utilizing open source labeled data to avoid the cost of manual labeling by legal experts, resulting in the best performance reported to date in this field. This work also opens avenues for future investigation into interpretable graph models and pre-trained models for long-text similarity training.

## ACKNOWLEDGMENT

The authors would like to thank the Key Laboratory of Sichuan Province on the Empirical Legal Studies and Smart Rule of Law, Sichuan University, for the provision of computational resources instrumental in conducting the research presented in this article.

## REFERENCES

[1] P. Bhattacharya, K. Hiware, S. Rajgaria, N. Pochhi, K. Ghosh, and S. Ghosh, "A comparative study of summarization algorithms applied to legal case judgments," in *Proc. 41st Eur. Conf. IR Res. (ECIR)*, in Lecture Notes in Computer Science, vol. 11437, L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra, Eds. Cologne, Germany: Springer, 2019, pp. 413–428, doi: 10.1007/978-3-030-15712-8_27.

[2] M.-Y. Kim and R. Goebel, "Two-step cascaded textual entailment for legal bar exam question answering," in *Proc. 16th, Ed., Int. Conf. Articial Intell. Law*, Jun. 2017, pp. 283–290.

[3] H. Chen, D. Cai, W. Dai, Z. Dai, and Y. Ding, "Charge-based prison term prediction with deep gating network," 2019, *arXiv:1908.11521*.

[4] B. Luo, Y. Feng, J. Xu, X. Zhang, and D. Zhao, "Learning to predict charges for criminal cases with legal basis," 2017, *arXiv:1707.09168*.

[5] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, "Large-scale multi-label text classification on EU legislation," 2019, *arXiv:1906.02192*.

[6] Y. Kano et al., "COLIEE-2018: Evaluation of the competition on legal information extraction and entailment," in *New Frontiers in Artificial Intelligence*, K. Kojima, M. Sakamoto, K. Mineshima, and K. Satoh, Eds. Cham, Switzerland: Springer, 2019, pp. 177–192.

[7] Q. Zhong, X. Fan, X. Luo, and F. Toni, "An explainable multi-attribute decision model based on argumentation," *Expert Syst. Appl.*, vol. 117, pp. 42–61, Mar. 2019.

[8] S. Althammer, A. Askari, S. Verberne, and A. Hanbury, "DoSSIER@COLIEE 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval," 2021, *arXiv:2108.03937*.

[9] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, T. Zhang, X. Han, Z. Hu, H. Wang, and J. Xu, "CAIL2019-SCM: A dataset of similar case matching in legal domain," 2019, *arXiv:1911.08962*.

[10] W. Yu, Z. Sun, J. Xu, Z. Dong, X. Chen, H. Xu, and J.-R. Wen, "Explainable legal case matching via inverse optimal transport-based rationale extraction," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2022, pp. 657–668.

[11] J. F. Horty, "Rules and reasons in the theory of precedent," *Legal Theory*, vol. 17, no. 1, pp. 1–33, Mar. 2011.

[12] C. Dent and I. Cook, "Stare decisis, repetition and understanding common law," *Griffith Law Rev.*, vol. 16, no. 1, pp. 131–150, Jan. 2007.

[13] Y. Ma, Y. Shao, Y. Wu, Y. Liu, R. Zhang, M. Zhang, and S. Ma, "LeCaRD: A legal case retrieval dataset for Chinese law system," in *Proc. 44th Int. ACM Sigir Conf. Res. Develop. Inf. Retr.*, 2021, pp. 2342–2348.

[14] M. van Opijnen and C. Santos, "On the concept of relevance in legal information retrieval," *Artif. Intell. Law*, vol. 25, no. 1, pp. 65–87, Mar. 2017.

[15] S. Khalid, S. Wu, A. Wahid, A. Alam, and I. Ullah, "An effective scholarly search by combining inverted indices and structured search with citation networks analysis," *IEEE Access*, vol. 9, pp. 120210–120226, 2021.

[16] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *J. Amer. Soc. Inf. Sci.*, vol. 27, no. 3, pp. 129–146, May 1976.

[17] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988.

[18] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.

[19] F. Song and W. B. Croft, "A general language model for information retrieval," in *Proc. 8th Int. Conf. Inf. Knowl. Manag.*, 1999, pp. 316–321.

[20] S. Clinchant and P. Florent, "Aggregating continuous word embeddings for information retrieval," in *Proc. Workshop Continuous Vector Space Models Their Compositionality*, 2013, pp. 100–109.

[21] D. Gillick, A. Presta, and G. Singh Tomar, "End-to-end retrieval in continuous space," 2018, *arXiv:1811.08008*.

[22] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, "Dense passage retrieval for open-domain question answering," 2020, *arXiv:2004.04906*.

[23] J. Zhan, J. Mao, Y. Liu, M. Zhang, and S. Ma, "RepBERT: Contextualized text embeddings for first-stage retrieval," 2020, *arXiv:2006.15498*.

[24] O. Khattab and M. Zaharia, "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2020, pp. 39–48.

[25] L. Gao, Z. Dai, and J. Callan, "COIL: Revisit exact lexical match in information retrieval with contextualized inverted list," 2021, *arXiv:2104.07186*.

[26] Y. Luan, J. Eisenstein, K. Toutanova, and M. Collins, "Sparse, dense, and attentional representations for text retrieval," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 329–345, Apr. 2021.

[27] S. Wehnert, S. A. Hoque, W. Fenske, and G. Saake, "Threshold-based retrieval and textual entailment detection on legal bar exam questions," 2019, *arXiv:1905.13350*.

[28] B. Lv and W.-L. Hou, "Typical case recommendation of court texts based on topic model," *Microelectron. Comput.*, vol. 35, no. 2, pp. 128–132, 2018.

[29] Y. Wang, J. Ge, Y. Zhou, Y. Feng, C. Li, Z. Li, X. Zhou, and B. Luo, "Topic model based text similarity measure for Chinese judgment document," in *Data Science: Third International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2017, Changsha, China, September 22–24, 2017, Proceedings, Part II*. Springer, 2017, pp. 42–54.

[30] W. Deng, "Research on judicial intelligence based on deep learning," M.S. thesis, Harbin Inst. Technol., Heilongjiang, China, 2017.

[31] L. Li, "Computing document similarity for the legal case retrieval," M.S. thesis, School Comput. Sci. Technol., Nanjing Normal Univ., Nanjing, China, 2018.

[32] S. Bano, S. Khalid, N. M. Tairan, H. Shah, and H. A. Khattak, "Summarization of scholarly articles using BERT and BiGRU: Deep learning-based extractive approach," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 9, Oct. 2023, Art. no. 101739.

[33] S. Bano and S. Khalid, "BERT-based extractive text summarization of scholarly articles: A novel architecture," in *Proc. Int. Conf. Artif. Intell. Things (ICAIoT)*, Dec. 2022, pp. 1–5.

[34] Y. Shao, J. Mao, Y. Liu, W. Ma, K. Satoh, M. Zhang, and S. Ma, "BERT-PLI: Modeling paragraph-level interactions for legal case retrieval," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 3501–3507.

[35] W. Hu, S. Zhao, Q. Zhao, H. Sun, X. Hu, R. Guo, Y. Li, Y. Cui, and L. Ma, "BERT_LF: A similar case retrieval method based on legal facts," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–9, Apr. 2022.

[36] J. Fang, X. Li, and Y. Liu, "Low-resource similar case matching in legal domain," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2022.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[38] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 7370–7377.

[39] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[40] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[41] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021.

[42] J. Rabelo, M. Kim, R. Goebel, M. Yoshioka, Y. Kano, and K. Satoh, "COLIEE 2020: Methods for legal document retrieval and entailment," in *Proc. New Frontiers Artif. Intell.-JSAI-isAI Workshops JURISIN (LENLS)*, in Lecture Notes in Computer Science, vol. 12758, N. Okazaki, K. Yada, K. Satoh, and K. Mineshima, Eds. Springer, 2020, pp. 196–210, doi: 10.1007/978-3-030-79942-7_13.

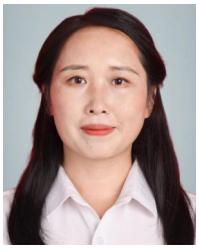
[43] G. Salton, E. A. Fox, and H. Wu, "Extended Boolean information retrieval," *Commun. ACM*, vol. 26, no. 11, pp. 1022–1036, Nov. 1983.

[44] C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun, "Lawformer: A pre-trained language model for Chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666651021000176

[45] S. Khalid, S. Wu, and F. Zhang, "A multi-objective approach to determining the usefulness of papers in academic search," *Data Technol. Appl.*, vol. 55, no. 5, pp. 734–748, Oct. 2021.

**YAO LIU** received the B.S. and M.S. degrees from the School of Computer Science, Southwest Petroleum University, China, in 2007. He is currently pursuing the Ph.D. degree with the School of Computer Science, Universiti Sains Malaysia, Penang, Malaysia. He is currently an Associate Professor with the Department of Management and Media, The Engineering and Technology College, Chengdu University of Technology. His research interests include natural language processing and legal artificial intelligence.

**TIEN-PING TAN** received the Ph.D. degree from Université Joseph Fourier, France, in 2008. He is currently an Associate Professor with the School of Computer Sciences, Universiti Sains Malaysia. His research interests include automatic speech recognition, machine translation, and natural language processing.



**XIAOPING ZHAN** received the Ph.D. degree from the Southwestern University of Finance and Economics, Statistics, in 2017. She is currently an Associate Professor with the School of Law, Sichuan University. She is also the Deputy Director of the Key Laboratory of Sichuan Province on the Empirical Legal Studies and Smart Rule of Law. Her research interests include procedural jurisprudence, data jurisprudence, and the legal regulation of privacy-enhancing technologies.