# Prompt-based Effective Input Reformulation for Legal Case Retrieval

Yanran Tang, Ruihong Qiu, and Xue Li

The University of Queensland {yanran.tang, r.qiu}@uq.edu.au, xueli@eecs.uq.edu.au

Abstract. Legal case retrieval plays an important role for legal practitioners to effectively retrieve relevant cases given a query case. Most existing neural legal case retrieval models directly encode the whole legal text of a case to generate a case representation, which is then utilised to conduct a nearest neighbour search for retrieval. Although these straightforward methods have achieved improvement over conventional statistical methods in retrieval accuracy, two significant challenges are identified in this paper: (1) Legal feature alignment: the usage of the whole case text as the input will generally incorporate redundant and noisy information because, from the legal perspective, the determining factor of relevant cases is the alignment of key legal features instead of whole text matching; (2) Legal context preservation: furthermore, since the existing text encoding models usually have an input length limit shorter than the case, the whole case text needs to be truncated or divided into paragraphs, which leads to the loss of the global context of legal information. In this paper, a novel legal case retrieval framework, PromptCase, is proposed to tackle these challenges. Firstly, *legal facts* and *legal issues* are identified and formally defined as the key features facilitating legal case retrieval based on a thorough study of the definition of relevant cases from a legal perspective. Secondly, with the determining legal features, a promptbased encoding scheme is designed to conduct an effective encoding with language models. Extensive zero-shot experiments have been conducted on two benchmark datasets in legal case retrieval, which demonstrate the superior retrieval effectiveness of the proposed PromptCase. The code has been released on https://github.com/yanran-tang/PromptCase.

Keywords: Legal case retrieval · Information retrieval

### 1 Introduction

Legal case retrieval (LCR) aims to retrieve relevant cases given a query case, which is important for legal practitioners in the world's two major legal systems, common law and civil law. From a legal perspective, the precedents are the historical cases that are similar to a given case in two determining aspects, legal facts and legal issues. In common law system, the judicial reasons of a judgement are critically based on the relevant cases, which is also called "the doctrine of precedents" [13]. While in civil law system, although the judgement is

not necessarily to be based on previously relevant cases, judges and lawyers are still strongly suggested to obtain legal information from these relevant cases <sup>1</sup>. Nowadays, the methods of LCR can be generally divided into two branches, statistical retrieval models [14,27,32] that measure the term frequency similarity between cases and neural LCR models [1,2,6,8,9,16,17,21,29,33,36–38,41] that encode the case into a representation to conduct nearest neighbour search.

Recently, neural LCR models have greatly attracted the research focus for the outstanding text representation ability. Generally, BERT-based legal case retrieval models use the whole text of a case to obtain the representation of the case, which directly determines the retrieval quality and accuracy with the case similarity calculation afterwards [1, 8, 9, 37, 38]. Due to the input length limit of BERT-based models, e.g., 512 tokens [11], a case is typically too long to be directly fed into these models with more than 5,000 words in a case. Therefore, most of these methods rely on truncating the case text to a suitable length [16] or dividing the whole text into smaller segments to process the input text [33].

Although these models have achieved competitive progress compared to traditional statistical models in LCR, there are still two challenges remaining: (1) Legal feature alignment: using the whole case text as the input for case representation generation [9,38] will incorporate redundant and noisy information, because from legal perspective, the determining factor of relevant cases is the alignment of legal features instead of whole text matching. When legal practitioners are retrieving relevant cases, they are actually finding "precedents", which refer to a court decision in an earlier case with legal facts and legal issues similar to the current case<sup>2</sup> Therefore, similar legal facts and legal issues are the key to retrieving relevant cases considering legal theory. (2) Legal context preservation: furthermore, the whole case text is usually truncated [16] or divided into paragraphs [33] due to the input length limit of BERT-based models, which is ineffective in capturing the legal context information. A legal case generally contains more than 5,000 words (in certain situations, easily exceeding 50,000 words), which is much longer than the 512-token input limit for BERT [11], 16k-token for Longformer [7], or 8k-token for ChatGPT [26]. Thus, passively truncating or dividing the case will lead to a significant loss of decisive legal features and case global view among the legal context information.

In light of the above observations, a novel LCR framework called PromptCase is proposed in this paper to tackle these challenges. Firstly, the input representation with two determining legal features, legal facts and legal issues, are proposed to effectively obtain representative legal information in cases instead of using the entire case. According to the formal legal document writing requirements, the format of a case text is well structured so that *legal facts* and *legal issues* can be effectively extracted from the case with sufficient processing steps. Secondly, in order to effectively encode the extracted legal features, a novel prompt-based encoding scheme is proposed to encode these features with language models. Empirical experiments are conducted on two benchmark datasets, LeCaRD [22]

<sup>&</sup>lt;sup>1</sup> https://www.court.gov.cn/zixun-xiangqing-243981.html

<sup>&</sup>lt;sup>2</sup> https://www.uscourts.gov/glossary

and COLIEE [12], which shows that the specific legal features proposed in this paper can represent the legal case more precisely to make a good representation for neural LCR models and effectively improve the performance of neural LCR models. The main contributions of this paper are summarised as follows:

- A PromptCase model is proposed for effective legal case retrieval by tackling the legal feature alignment and legal context preservation challenges.
- Two determining legal features, *legal facts* and *legal issues* are identified and extracted from legal cases with adequate processing procedures.
- A prompt-based encoding scheme is derived to effectively encode the extracted legal features for the widely used language models.
- Extensive experiments conducted on two benchmark datasets demonstrate the state-of-the-art performance of the PromptCase framework.

## 2 Related Work

### 2.1 Legal Case Retrieval

LCR is a special type of IR. The methods of IR can be generally divided into two branches, statistical methods [14,27,32] and neural network methods [15,25,28,31]. Similarly, in LCR there are the same two branches. Statistical models include TF-IDF [14], BM25 [32] and LMIR [27], which rely on the term frequency and inverse document frequency of words to determine the similarity between cases. Neural LCR models rely on encoding the case using the language models [10, 11, 20, 24]. With the increasing amount of online legal information and users' legal information needs, many neural LCR models [1,2,6,8,9,16,17,21,29,33,34,36-38,40] are conducted to bridge the information gap by capturing domain-specific and personal needs. Law2Vec [9] is a legal language model that pre-trains on a large legal corpus. Lawformer [38] focuses on combining three types of attention mechanisms to get the context of long legal cases. BERT-PLI [33] calculates the similarity between two paragraphs of cases text to tackle the lengthy problem of legal cases. SAILER [16] is a pre-trained language model that selects the reasoning, decision and fact sections in the cases to train the encoder and uses the fact section to be the input of the encoder to get the case representation.

### 2.2 Input Reformulation in Neural Legal Case Retrieval

Input reformulation plays an important role in neural LCR because a case is hard to fit into the model directly due to the length limit [3–5, 16, 35, 39]. Askari et al. [5] and LeiBi [4] both propose to combine lexical and neural network methods to get a summary of a legal case as the case representation. LEVEN [39] utilises the frequency of legal events to reformulate the case input. Both CL4LJP [41] and QAjudge [43] intuitively reformulate the case input with only the fact instead of the whole case. IOT-Match [40] reformulates the case input based on legal rationales. BERT-PLI [33] divides the case input into the paragraph-level interaction between query and candidate cases. Liu et al. [18, 19] proposes to use the conversational search paradigm to reformulate the query case.

### 3 Preliminary

### 3.1 Task Definition

In legal case retrieval, given a query case q, and a set of n candidate cases, denoted as  $\mathcal{D} = \{d_1, d_2, ..., d_n\}$ , the task is to retrieve a set of relevant cases  $\mathcal{D}^* = \{d_i^* | d_i^* \in \mathcal{D} \land relevant(d_i^*, q)\}$  from  $\mathcal{D}$ , where  $relevant(d_i^*, q)$  denotes that  $d_i^*$  is a relevant case of the query case q. From a legal perspective, the relevant cases are called precedents, which are the historical cases with legal facts and legal issues similar to the given query case. Specifically, given a query case, the relevant cases in COLIEE2023 dataset are the cases referred by the query case. While in LeCaRD dataset, cases having similar key facts and key circumstances to the query case are labelled as relevant cases by legal experts.

#### 3.2 Input Reformulation in Neural Legal Case Retrieval

Existing neural LCR models generally use the full case as the input to the model with different input reformulation methods to deal with the overly long cases.

*BERT-PLI* [33] reformulates the case input into the paragraph-level interaction vector between the query and candidate cases as below:

$$\mathbf{e}_{(a_i,d_i)} = \text{BERT}([\text{CLS}]; q_i; [\text{SEP}]; d_j; [\text{SEP}]), \tag{1}$$

where ";" denotes the concatenation function, and [CLS] and [SEP] are two special tokens for BERT to denote the input's beginning and separation.  $q_i$  and  $d_j$  are the *i*-th paragraph and *j*-th paragraph of case q and d.

SAILER [16] uses the fact section of cases as the input of a finetuned BERT:

$$\mathbf{e}_q = \text{BERT}([\text{CLS}]; q_{(\text{fact})}; [\text{SEP}]), \quad \mathbf{e}_d = \text{BERT}([\text{CLS}]; d_{(\text{fact})}; [\text{SEP}]), \quad (2)$$

where  $q_{(\text{fact})}$  and  $d_{(\text{fact})}$  are the fact of the query case and the document case respectively. The fact is assumed to be located in the most front and if the length of the fact is longer than 512 tokens, the first 512 tokens of the case will be used.

*BM25Inject* [3] concatenates the BM25 score of the query case and the document case into the input of BERT-based cross-encoder:

$$s_{(q,d)} = \operatorname{BERT}([\operatorname{CLS}]; q; [\operatorname{SEP}]; s_{\operatorname{BM25}_{(q,d)}}; [\operatorname{SEP}]; d; [\operatorname{SEP}]),$$
(3)

where  $s_{BM25_{(q,d)}}$  is the BM25 score scalar of the query case q and the candidate d and the final semantic similarity is  $s_{(q,d)}$ .

### 4 Method

In this section, the PromptCase framework will be introduced. In Section 4.1, two determining legal features are extracted. A prompt-based method utilising these legal features will be detailed in Section 4.2. The measurement of case similarity will be introduced in Section 4.3. The overview of PromptCase is shown in Fig. 1.



Prompt-based Effective Input Reformulation for Legal Case Retrieval

Fig. 1: The framework of PromptCase. LM means a language model, e.g., BERT. The final output of the [CLS] token is the representation embedding of a legal fact, a legal issue or a case. (a) The process of legal facts and legal issues extraction. When legal facts are not explicitly available, ChatGPT is applied to generate a case summary as legal facts. (b) Dual and cross encoding with prompt of a case.

Lafond v. Muskeg Lake Cree Na3on (2008), 330 F.T.R. 60 (FC)	李月航容留他人吸毒一案 <mark>(Case name)</mark>
Background On February 13, 2006, the applicant was elected as a councillor to the MLCN Band Council for a term of three years. The respondent Band is located in the province of Saskatchewan	<b>案件基本情况 (Background)</b> 长乐市人民检察院指控: 1、2017年9月25日22时许,被告人李月航 在其租住的长乐市某街道某村某公寓房间内,容留王某吸食甲基苯丙胺 (俗称"冰毒")。 2、2017年10月19日晚,被告人李月航在其租住的 长乐市某街道某村某公寓房间内,容留王某 <b>经审理查明:</b> 1、2017年
Analysis	9 月 25 日 22 时许,被告人李月航在其租住的长
Does this Court have jurisdiction over the present application? In order to determine the jurisdiction of the Federal Court in this matter, it is imperative to Indeed this was recognized by the Federal Court of Appeal in FRAGMENT_SUPPRESSED, where it held that FRAGMENT_SUPPRESSED. I agree that the Chief does have inherent	裁判分析过程 (Analysis) 本院认为,被告人李月航多次为他人吸食毒品提供场所,其行为已构成容 留他人吸毒罪。长乐市人民检察院指控的罪名成立,应依法追究被告人李 月航的刑事责任。被告人李月航因涉嫌吸毒被公安机关抓获,主动向公安 机关供述了尚未被掌握的其容留他人吸毒的犯罪事实,视为自动投案,系 自首,依法可从轻处罚;被告人李月航被公安
Order For these reasons, the application for judicial review of Chief Ledoux's decision will be allowed. (a) COLIEE dataset (common law)	<b>判決結果 (Judgement)</b> 被告人李月航犯容留他人吸毒罪,判处拘役五个月,并处罚金人民币三千 元。 (b) LeCaRD dataset (civil law)

Fig. 2: Example of case documents

#### 4.1**Extraction of Legal Facts and Legal Issues**

This section describes the extraction of legal facts and legal issues from cases as shown in Fig. 1(a) to overcome the legal feature alignment challenge. For common law (COLIEE dataset) or civil law (LeCaRD dataset) respectively, a case often has a relatively fixed writing style, which includes four basic parts as in Fig. 2. The first part is the case name with basic information about the case. The second part is the "Background" of the case demonstrating detailed information about the case. The third part is "Analysis" describing the reasons why the judges make the final decision. The final part called "Order" or "Judgement", is the judgement of the case. Such a clear and general structure of legal cases provides access to locate and extract legal facts and legal issues from extremely long cases.

Legal facts. Legal fact is a fundamental part that describes the "who, when, what, where and why" in legal cases. Firstly, in the COLIEE2023 dataset, the

detailed process of a case is generally written in the background part, which is often more than thousands of words that will exceed the input limit of BERT-based models. In order to get an abstract yet accurate legal facts of cases, the ChatGPT [26] is used to get the summary of legal facts. The ChatGPT API with "gpt-3.5-turbo" model is used with the prompt of "Summarise in 50 words: ". As a result, the output of ChatGPT will be the legal facts  $c_{\text{(fact)}}$  of the case c.

Secondly, in LeCaRD, the fact section is a separate and brief part that can be found in "Background", beginning with a description of "After the trial, it was found out that: " in Chinese (the bold Chinese words "经审理查明: " in the "Background" part in Fig. 2(b)). Thus, in LeCaRD, the legal facts  $c_{\text{(fact)}}$  of the case c are extracted directly based on the understanding of a legal case.

Legal issues. The definition of "issue" in legal domain is "a critical feature that focuses on the dispute points between the parties in the case."<sup>3</sup> In case documents of common law, the legal issues are located in the "Analysis" part, which is given by the judges to settle the disputes between the parties with legal reasons. To have convincing reasons, the judges will list the relevant precedents' facts, issues or judgements in this part to support the judges' opinions. Specifically, as shown in Fig. 2(a), there are words replaced by placeholders with special terms in cases of the COLIEE2023 dataset, such as "FRAGMENT\_SUPPRESSED". The original words for these placeholders are the case name of a precedent. These placeholders are for the task of legal case retrieval, which is to find the precedents being referred in the placeholder. Thus, for the COLIEE dataset, all of the sentences with placeholders will be selected as the legal issues  $c_{(issue)}$  of the case c.

Compared to common law, the judges in the civil law system often make their judgements according to the legal articles written in the acts while the judges of the common law system have the compulsory responsibility to refer the precedents to support their final decisions. And there is also no specific part for settling legal issues in the cases of civil law. After a thorough study of the cases from LeCaRD dataset under the civil law system, it is found that legal issues often appear in the case as the name of charges, such as "murder". Therefore, the names of charges in Chinese criminal law are collected and saved as a list of charges. For every case (queries and candidates) in LeCaRD, the full text of a case will be used to find the charges that appear both in the case and the list of charges. Finally, all of the found charges are the legal issues  $c_{(issue)}$  of the case c.

### 4.2 Prompt-based Case Encoding

After extracting legal facts and legal issues, a prompt-based case encoding method is developed in this section to tackle the legal context preservation challenge.

**Prompt template.** With the recent advances of prompt, the capability of prompting a language model is impressive in understanding the context information of a task. To enable the language models to capture the global context

<sup>&</sup>lt;sup>3</sup> https://www.uscourts.gov/glossary

of legal information, the prompt templates of "Legal facts:" ("法律事实:" in Chinese) and "Legal issues:" ("法律纠纷:" in Chinese) will be added to the beginning of the legal facts and legal issues texts and fed into the language model together. For every legal case in COLIEE2023 and LeCaRD datasets, the prompt template is formulated as below:

$$prompt_{(fact)} = "Legal facts:", prompt_{(issue)} = "Legal issues:".$$
 (4)

**Dual encoding with prompt.** To avoid the undesired cross-effect between legal facts and legal issues, the legal facts with prompt and legal issues with prompt will be fed into the BERT-based encoder separately to get the individual legal facts embedding  $\mathbf{e}_{\text{dual},c_{(\text{fact})}}$  and legal issues embedding  $\mathbf{e}_{\text{dual},c_{(\text{issue})}}$ . The encoding process can be denoted as the following equations:

$$\mathbf{e}_{\text{dual},c_{(\text{fact})}} = \text{LM}([\text{CLS}]; \text{prompt}_{(\text{fact})}; c_{(\text{fact})}; [\text{SEP}]), \\
\mathbf{e}_{\text{dual},c_{(\text{issue})}} = \text{LM}([\text{CLS}]; \text{prompt}_{(\text{issue})}; c_{(\text{issue})}; [\text{SEP}]),$$
(5)

where  $\mathbf{e}_{\text{dual},c_{(\text{fact})}}$  and  $\mathbf{e}_{\text{dual},c_{(\text{issue})}}$  are both the embedding of the final hidden state of the [CLS] token of the language model (LM), e.g., BERT.

**Cross encoding with prompt.** On the contrary, to obtain the deeper interactions between legal facts and legal issues, the cross encoding method is also being conducted as the following equations:

$$\mathbf{e}_{\text{cross},c} = \text{LM}([\text{CLS}]; \text{prompt}_{(\text{fact})}; c_{(\text{fact})}; [\text{SEP}]; \text{prompt}_{(\text{issue})}; c_{(\text{issue})}; [\text{SEP}]).$$
(6)

where  $\mathbf{e}_{\text{cross},c}$  is the output embedding of the [CLS] token of LM.

**Case representation** To obtain both the original and interaction information of legal facts and legal issues, the case representation will be the concatenation of the  $\mathbf{e}_{\text{dual},c_{\text{(fact)}}}$ ,  $\mathbf{e}_{\text{dual},c_{\text{(issue)}}}$ , and  $\mathbf{e}_{\text{cross},c}$  as the following equations:

$$\mathbf{e}_{c} = \mathbf{e}_{\mathrm{dual},c_{(\mathrm{fact})}}; \mathbf{e}_{\mathrm{dual},c_{(\mathrm{issue})}}; \mathbf{e}_{\mathrm{cross},c}.$$
(7)

#### 4.3 Case Similarity

Similar to traditional IR tasks, the dot product (denoted as  $(\cdot)$ ) is used to measure the semantic similarity between two cases. Given the case representation  $\mathbf{e}_q$  and  $\mathbf{e}_d$  of case q and candidate case d generated by PromptCase, the similarity score  $s_{(q,d)}$  is calculated as:

$$s_{(q,d)} = \mathbf{e}_q \cdot \mathbf{e}_d. \tag{8}$$

### 5 Experiments

5.1 Setup

**Datasets.** To evaluate the proposed PromptCase, the experiments are conducted on the following LCR datasets with summarised statistics in Table 1.

**LeCaRD** [22]. LeCaRD is a legal case retrieval dataset, where the cases are from the supreme court of China, a civil law system country. It contains 107 queries and over 43,000 candidate cases. For each cuery, there is a candidate pool of 100 case

Table 1:	Statistics	of	LeCaRD	and
COLIEE	2023  datas	set	5.	

Datasets	LeCaRD COLIEE2023						
Language	Chinese	English					
Avg. length/case	8,275	5,566					
Largest length of cases	99,163	61,965					
Avg. relevant cases/query	10.33	2.69					

query, there is a candidate pool of 100 cases. The evaluation of LeCaRD is based on the binary golden label for a more restrict requirement  $^4$ .

**COLIEE2023** [12] <sup>5</sup>. COLIEE2023 is a dataset from Competition on Legal Information Extraction/Entailment (COLIEE) 2023, where cases are from the federal court of Canada with common law system. Given a query case, relevant cases are retrieved from the entire candidate pool. To avoid the data leakage problem of pre-trained models, only the testing set of COLIEE2023 is used.

Metrics. For both datasets, precision (P), recall (R), Micro F1 (Mi-F1), Macro F1 (Ma-F1), Mean Reciprocal Rank (MRR), Mean Average Precision (MAP) and normalized discounted cumulative gain (NDCG) are used. For both LeCaRD and COLIEE2023 datasets, top 5 ranking results are evaluated by following previous methods [12, 16, 22]. All metrics are the higher the better.

**Baselines.** The following baselines are chosen for comparison:

- BM25 [32] is a statistical retrieval model using the term frequency and inverse document frequency, which is still a strong baseline.
- BERT [11] is a strong bi-directional transformer encoder in language tasks.
   For LeCaRD in Chinese, the "uer/sbert-base-chinese-nli" [42] model is used , while for COLIEE2023 in English, the "bert-base-uncased" [11] model is used.
- Lawformer [38] is pre-trained on Chinese legal corpus and focuses on long documents processing.
- LEGAL-BERT [8] is pre-trained on a large English legal corpus and achieves state-of-the-art results in different legal understanding tasks.
- MonoT5 [24] is a pre-trained sequence-to-sequence model focuses on document ranking task using the powerful T5 model [30].
- SAILER [16] is a structure-aware pre-trained model that achieves stateof-the-art performance on both datasets. Two-stage usage of SAILER with BM25 is evaluated as well.

BERT-PLI [33] is not compared since its paragraph-level interaction is not applicable to legal facts and legal issues. BM25Inject [3] is not compared because its cross encoding between cases is not extendable in our scenario.

<sup>&</sup>lt;sup>4</sup> https://github.com/myx666/LeCaRD#golden\_labelsjson

<sup>&</sup>lt;sup>5</sup> https://sites.ualberta.ca/~rabelo/COLIEE2023/

**Implementation.** The French text in COLIEE2023 is removed. The two-stage method is based on the top 10 retrieved cases by BM25 model. All experiments are in a zero-shot manner without training, except that the SAILER model for COLIEE2023 is pre-trained on the COLIEE2023 training set. The experiment of BM25 model with PromptCase reformulated input utilises the original text, legal facts, legal issues and prompt together.

### 5.2 Overall Performance

In this section, the PromptCase is evaluated by being integrated into the baselines. The results are presented in Table 2 for LeCaRD and Table 3 for COLIEE2023.

Overall, the PromptCase can steadily improve the performances of all baselines by a large margin. With the state-of-the-art pretrained SAILER model in legal domain, PromptCase significantly boosts the retrieval performance for both one and two stage manners with a proper reformulation of case input. For the traditional method BM25, the performance of using PromptCase is better than with the whole case as input. The improved performance shows that the reformulated input can capture the determining legal features with proper emphasis on the term frequency without being biased by the long and noisy case texts. For the pretrained BERT with full case as input, the performances on both datasets are worse than BM25 and SAILER. However, BERT+PromptCase outperforms BM25+PromptCase and the SAILER baseline model on LeCaRD, which in-

Table	2:	Overall	performance	on	LeCaRD	(%)	
-------	----	---------	-------------	----	--------	-----	--

Methods	LeCaRD@5										
Methods	P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5				
BM25	40.0	19.2	26.0	30.5	58.3	48.5	45.9				
+ PromptCase	41.3	19.9	26.8	31.7	60.6	58.8	65.2				
BERT	38.7	18.6	25.1	26.7	57.4	54.3	61.0				
+ PromptCase	46.2	22.2	30.0	35.4	64.4	61.2	67.9				
Lawformer	29.0	13.9	18.8	19.5	43.6	41.9	48.2				
+ PromptCase	38.9	18.7	25.3	30.7	62.0	59.7	64.0				
SAILER	46.7	22.5	30.4	37.1	67.9	65.4	70.1				
+ PromptCase	51.6	24.8	33.5	43.0	71.1	67.6	74.2				
Two-stage											
SAILER	47.8	23.0	31.1	36.1	67.3	64.4	70.6				
+PromptCase	51.0	24.6	33.2	38.7	70.7	67.9	73.5				
						0 T TT					

Table 3: Overall performance on COLIEE (%).

Methods	COLIEE2023										
internetab	P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5				
BM25 +PromptCase	$\begin{vmatrix} 16.5 \\ 17.0 \end{vmatrix}$	$\begin{array}{c} 30.6\\ 31.5 \end{array}$	$21.4 \\ 22.1$	$22.2 \\ 23.0$	$23.1 \\ 24.2$	$\begin{array}{c} 20.4\\ 21.6 \end{array}$	$23.7 \\ 24.4$				
BERT +PromptCase	$\begin{vmatrix} 2.07 \\ 2.38 \end{vmatrix}$	$\begin{array}{c} 3.84 \\ 4.42 \end{array}$	$2.69 \\ 3.10$	$2.57 \\ 3.02$	$5.51 \\ 6.33$	$\begin{array}{c} 5.48\\ 6.25\end{array}$	6.25 7.21				
LEGAL-BERT +PromptCase	$\begin{array}{c} 4.64 \\ 4.83 \end{array}$	$\begin{array}{c} 8.61 \\ 8.96 \end{array}$	$6.03 \\ 6.28$	$6.03 \\ 6.44$	$11.4 \\ 13.4$	$\begin{array}{c} 11.3\\ 13.4 \end{array}$	$13.6 \\ 15.5$				
MonoT5 +PromptCase	$\begin{array}{c} 0.38 \\ 0.56 \end{array}$	$\begin{array}{c} 0.70 \\ 1.05 \end{array}$	$0.49 \\ 0.73$	$0.47 \\ 0.72$	$1.17 \\ 1.63$	$\begin{array}{c} 1.33 \\ 1.43 \end{array}$	$0.61 \\ 0.89$				
SAILER +PromptCase	12.8 16.0	$\begin{array}{c} 23.7\\ 29.7\end{array}$	$16.6 \\ 20.8$	$17.0 \\ 21.5$	$25.9 \\ 32.7$	$25.3 \\ 32.0$	29.3 36.2				
Two-stage SAILER +PromptCase	19.6 21.8	32.6 36.3	24.5 27.2	$23.5 \\ 26.5$	37.3 39.9	36.1 38.7	40.8 44.0				

dicates that BERT is a semantic LM that can better understand and represent a case using legal features semantics. While the term frequency cannot fully take advantage of the semantics in legal facts and legal issues, which also limits the performance of two-stage SAILER on LeCaRD with or without PromptCase

								-				-		· ·		
Prompt	ompt Leg-Fea					LeC	aRD	COLIEE2023								
			P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5	P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5
X	.	x	46.7	22.5	30.4	37.1	67.9	65.4	70.1	12.8	23.7	16.6	17.0	25.9	25.3	29.3
1	.	x	46.5	22.4	30.2	36.9	68.6	65.8	70.5	12.8	23.7	16.6	17.0	25.4	24.8	28.5
X	,	/	52.0	25.0	33.8	43.3	69.4	66.2	72.9	15.9	29.5	20.6	21.3	32.6	31.5	35.8
1	,	/	51.6	24.8	33.5	43.0	71.1	67.6	74.2	16.0	29.7	20.8	21.5	32.7	32.0	36.2

Table 4: Ablation study. Leg-Feat denotes legal features. (%)

compared with one-stage SAILER. Lawformer and LEGAL-BERT are two neural LCR models pre-trained on Chinese and English respectively, whose performances are improved significantly with PromptCase. The performance of MonoT5 is the worst in COLIEE2023 dataset, possibly because MonoT5 is pre-trained for text-to-text tasks different from retrieval tasks. Comparing the results on these two datasets, the improvement with PromptCase on LeCaRD is more obvious than on COLIEE2023. The possible reason is the different definitions of relevance in these datasets. For LeCaRD, the relevant cases are defined by legal experts, which is easier for models to identify. While for COLIEE2023, the relevant cases are referred cases by the query case, which are a subset of all relevant cases and not a golden label for relevance, leading to an inferior performance.

#### 5.3 Ablation Study

The ablation study is conducted to verify the effectiveness of the two main components of PromptCase, the legal features and the prompt encoding scheme. The SAILER [16] model is used as the base model in these experiments since SAILER is a state-of-the-art pre-trained model with English and Chinese on both datasets. Specifically, the prompt templates of the experiment without legal features are reformulated as "Legal facts and legal issues:" in English and "法律 事实和法律纠纷:" in Chinese for COLIEE2023 and LeCaRD respectively.

As shown in Table 4, the reformulated input with prompt and legal features can significantly improve the performance compared with other variants for both datasets. The legal features alone can largely increase the retrieval performance. While only using prompt encoding, the performance is not improved since there is no specific legal feature used with the prompt.

#### 5.4 Effectiveness of Legal Features

To verify the effectiveness of legal features, experiments are conducted using SAILER with: no legal features, only legal facts, only legal issues and both legal facts and legal issues. For no legal features, the second result in Table 4 is reused.

As shown in Table 5, the reformulated input with both legal facts and legal issues achieves the best performance in the effectiveness of legal features experiments, which indicates the challenge of legal feature alignment is well resolved. For LeCaRD, the performance of only using legal facts is better than

Facts	Issues			LeC	CaRD		COLIEE2023							
	P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5	P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5
x   x	43.0	20.7	27.9	34.0	63.5	51.7	51.7	12.8	23.7	16.6	17.0	25.4	24.8	28.5
✓   X	47.3	22.8	30.7	37.4	66.8	63.6	70.1	12.7	23.5	16.5	17.2	24.7	24.3	27.7
x   ✓	41.3	19.9	26.8	32.9	57.6	54.7	61.7	13.4	24.8	17.4	17.8	29.1	28.3	31.8
///	51.6	24.8	33.5	43.0	71.1	67.6	74.2	16.0	29.7	20.8	21.5	32.7	32.0	36.2

Table 5: Effectiveness of legal facts (Facts) and legal issues (Issues). (%)

Table 6: Effectiveness of different prompts. Instructive (IT): A: "Legal facts:/Legal issues:"; B: "The following is legal facts:/The following is legal issues:"; C: "The judge think:"; Misleading (ML): D: "This case is related to \$<randomly sample one issue>:"; E: "Legal facts of this case is \$<randomly sample one issue>:/Legal issues of this case is \$<randomly sample one issue>:"; G: "ADC is a database conference:" and NA: no prompt is used.

		LeCaRD							COLIEE2023							
		P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5	5 P@5	R@5	Mi-F1	Ma-F1	MRR@5	MAP	NDCG@5	
_	NA	52.0	25.0	33.8	43.3	69.4	66.2	72.9	15.9	29.5	20.6	21.3	32.6	31.5	35.8	
IT	A B C	51.6 51.4 51.8	24.8 24.7 24.9	33.5 33.3 33.6	43.0 42.7 43.5	71.1 71.0 70.1	$\begin{array}{c} 67.6 \\ 67.4 \\ 67.9 \end{array}$	74.2 74.0 74.1	$  \begin{array}{c} 16.0 \\ 15.9 \\ 15.7 \end{array}  $	$29.7 \\ 29.5 \\ 29.1$	20.8 20.6 20.4	21.5 21.4 21.1	32.7 32.8 32.0	32.0 32.0 31.3	36.2 36.0 35.8	
ML	D E	$\begin{vmatrix} 42.8 \\ 42.6 \end{vmatrix}$	$\begin{array}{c} 20.6 \\ 20.5 \end{array}$	$27.8 \\ 27.7$	30.8 29.7	$58.1 \\ 60.0$	$56.0 \\ 56.8$	$62.7 \\ 62.7$	$\begin{vmatrix} 14.5 \\ 15.1 \end{vmatrix}$	$\begin{array}{c} 26.9 \\ 28.1 \end{array}$	$\begin{array}{c} 18.8\\ 19.6\end{array}$	$\begin{array}{c} 19.6 \\ 20.5 \end{array}$	$28.7 \\ 29.5$	$\begin{array}{c} 27.8\\ 29.0 \end{array}$	31.9 33.4	
IR	F G	$51.4 \\ 51.6$	$\begin{array}{c} 24.7\\ 24.8\end{array}$	$33.4 \\ 33.5$	$\begin{array}{c} 42.9\\ 42.6\end{array}$	$69.3 \\ 69.9$	$\begin{array}{c} 66.5\\ 67.6\end{array}$	72.9 73.7	$\begin{vmatrix} 15.6 \\ 15.2 \end{vmatrix}$	$\begin{array}{c} 29.0\\ 28.2 \end{array}$	$\begin{array}{c} 20.3 \\ 19.7 \end{array}$	$21.1 \\ 20.5$	$32.4 \\ 32.1$	$\begin{array}{c} 31.4\\ 31.2 \end{array}$	$35.8 \\ 35.3$	

only using legal issues, while it is opposite in COLIEE2023 that only using legal issues is better than only using legal facts. This opposite phenomenon also appears in the experiments of ablation study. The different performances of datasets could be due to the different case structures in different legal systems, which may cause the different focuses of prompt and legal features.

#### 5.5 Effectiveness of Prompt

In this experiment, the effectiveness of Prompt is investigated with different prompt templates using SAILER. The prompt templates are widely chosen from instructive, misleading and irrelevant categories, which are detailed in Table 6.

As shown in Table 6, seven different prompt templates are selected to evaluate the effectiveness of prompts, which can be classified into three categories: instructive (correct legal prompts), misleading (wrongful legal prompts) and irrelevant (correct non-legal prompts). The performances of the experiments indicate that instructive prompts can improve performance by giving correct and informative indications of the global view of legal context. On the contrary, misleading prompts negatively impact the case retrieval accuracy. Compared with



Fig. 3: Visulisation of case encodings with and without PromptCase for LeCaRD.

the other categories of prompts, irrelevant prompts slightly hurts the performance by adding irrelevant noisy information to the input.

#### 5.6 Visualisation Analysis

To further prove the effectiveness of PromptCase input reformulation method, t-SNE [23] is used to visualise cases embeddings with and without PromptCase. Cases from five legally similar and difficult to distinguish charges of LeCaRD are selected to visualise in Fig. 3, including *theft*, *robbery*, *defraud*, *vandalism*, and *encroachment*. All selected case embeddings are generated by the zero-shot SAILER model. As shown in Fig. 3(a), case embeddings generated by SAILER are classified into three clusters. Moreover, vandalism cases are wrongfully classified as robbery cases and encroachment cases are wrongfully classified as defraud cases. Compared with SAILER, adding PromptCase (as shown in Fig. 3(b)) makes cases embeddings evenly distributed as five clusters corresponding to five charges, which indicates the powerful discriminative ability and the ability to learn legal context information of PromptCase framework.

### 6 Conclusion

This paper identifies the challenges in the existing LCR models about legal feature alignment and legal context preservation. To tackle these challenges, a novel legal case retrieval framework called PromptCase is introduced. In PromptCase, *legal facts* and *legal issues* are effectively extracted from the original case, which is further encoded with a prompt-based schema to generate an informative case representation. Extensive experiments are conducted on two benchmark datasets, which successfully demonstrate the superiority of PromptCase by achieving the best performance compared with state-of-the-art baselines.

**Acknowledgements** The work is supported by Australian Research Council CE200100025.

### References

- 1. Abolghasemi, A., Verberne, S., Azzopardi, L.: Improving bert-based query-bydocument retrieval with multi-task optimization. In: ECIR (2022)
- Althammer, S., Askari, A., Verberne, S., Hanbury, A.: Dossier@coliee 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval. CoRR abs/2108.03937 (2021)
- Askari, A., Abolghasemi, A., Pasi, G., Kraaij, W., Verberne, S.: Injecting the BM25 score as text improves bert-based re-rankers. In: ECIR (2023)
- Askari, A., Peikos, G., Pasi, G., Verberne, S.: Leibi@coliee 2022: Aggregating tuned lexical models with a cluster-driven bert-based model for case law retrieval. CoRR abs/2205.13351 (2022)
- 5. Askari, A., Verberne, S.: Combining lexical and neural retrieval with longformerbased summarization for effective case law retrieval. In: DESIRES. CEUR (2021)
- Askari, A., Verberne, S., Abolghasemi, A., Kraaij, W., Pasi, G.: Retrieval for extremely long queries and documents with RPRS: a highly efficient and effective transformer-based re-ranker. CoRR abs/2303.01200 (2023)
- Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. CoRR abs/2004.05150 (2020)
- Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: the muppets straight out of law school. CoRR abs/2010.02559 (2020)
- Chalkidis, I., Kampas, D.: Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artif. Intell. Law 27(2), 171–198 (2019)
- Dai, Z., Callan, J.: Context-aware sentence/passage term importance estimation for first stage retrieval. CoRR abs/1910.10687 (2019)
- 11. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019)
- Goebel, R., Kano, Y., Kim, M.Y., Loro, M.N., Minh, N.L., Rabelo, J., Rossi, J., Satoh, K., Savelka, J., Shao, Y., Shimazu, A., Tojo, S., Tran, V., Valvoda, J., Westermann, H., Yamada, H., Yoshioka, M., Wehnert, S.: Competition on legal information extraction/entailment (COLIEE) (2023)
- HARRIS, B.: Final appellate courts overruling their own "wrong" precedents: the ongoing search for principle. LAW QUARTERLY REVIEW 118(7), 408–427 (2002)
- 14. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. J. Documentation **60**(5), 493–502 (2004)
- 15. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In: SIGIR (2020)
- Li, H., Ai, Q., Chen, J., Dong, Q., Wu, Y., Liu, Y., Chen, C., Tian, Q.: SAILER: structure-aware pre-trained language model for legal case retrieval. CoRR abs/2304.11370 (2023)
- 17. Liu, B., Hu, Y., Wu, Y., Liu, Y., Zhang, F., Li, C., Zhang, M., Ma, S., Shen, W.: Investigating conversational agent action in legal case retrieval. In: ECIR (2023)
- Liu, B., Hu, Y., Wu, Y., Liu, Y., Zhang, F., Li, C., Zhang, M., Ma, S., Shen, W.: Investigating conversational agent action in legal case retrieval. In: ECIR (2023)
- Liu, B., Wu, Y., Zhang, F., Liu, Y., Wang, Z., Li, C., Zhang, M., Ma, S.: Query generation and buffer mechanism: Towards a better conversational agent for legal case retrieval. Inf. Process. Manag. (2022)
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019)

- 14 Y. Tang et al.
- Ma, Y., Ai, Q., Wu, Y., Shao, Y., Liu, Y., Zhang, M., Ma, S.: Incorporating retrieval information into the truncation of ranking lists for better legal search. In: SIGIR (2022)
- 22. Ma, Y., Shao, Y., Wu, Y., Liu, Y., Zhang, R., Zhang, M., Ma, S.: Lecard: A legal case retrieval dataset for chinese law system. In: SIGIR (2021)
- 23. van der Maaten, L., Hinton, G.: Viualizing data using t-sne. Journal of Machine Learning Research (2008)
- 24. Nogueira, R., Jiang, Z., Pradeep, R., Lin, J.: Document ranking with a pretrained sequence-to-sequence model. In: EMNLP (2020)
- Nogueira, R.F., Yang, W., Lin, J., Cho, K.: Document expansion by query prediction. CoRR abs/1904.08375 (2019)
- 26. OpenAI: Gpt-3.5-turbo (2021), https://openai.com/
- Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. SIGIR (2017)
- Qiao, Y., Xiong, C., Liu, Z., Liu, Z.: Understanding the behaviors of BERT in ranking. CoRR abs/1904.07531 (2019)
- Rabelo, J., Kim, M., Goebel, R.: Semantic-based classification of relevant case law. In: JURISIN (2022)
- 30. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. (2020)
- Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: EMNLP-IJCNLP (2019)
- 32. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: SIGIR (1994)
- Shao, Y., Mao, J., Liu, Y., Ma, W., Satoh, K., Zhang, M., Ma, S.: BERT-PLI: modeling paragraph-level interactions for legal case retrieval. In: IJCAI (2020)
- Sun, Z., Xu, J., Zhang, X., Dong, Z., Wen, J.: Law article-enhanced legal case matching: a model-agnostic causal learning approach. CoRR abs/2210.11012 (2022)
- Tran, V.D., Nguyen, M.L., Satoh, K.: Building legal case retrieval systems with lexical matching and summarization using A pre-trained phrase scoring model. In: ICAIL (2019)
- Vuong, T., Nguyen, H., Nguyen, T., Nguyen, H., Nguyen, T., Nguyen, H.: NOWJ at COLIEE 2023 - multi-task and ensemble approaches in legal information processing. CoRR abs/2306.04903 (2023)
- 37. Wang, Z.: Legal element-oriented modeling with multi-view contrastive learning for legal case retrieval. In: IJCNN (2022)
- Xiao, C., Hu, X., Liu, Z., Tu, C., Sun, M.: Lawformer: A pre-trained language model for chinese legal long documents. AI Open 2, 79–84 (2021)
- Yao, F., Xiao, C., Wang, X., Liu, Z., Hou, L., Tu, C., Li, J., Liu, Y., Shen, W., Sun, M.: LEVEN: A large-scale chinese legal event detection dataset. In: ACL (2022)
- Yu, W., Sun, Z., Xu, J., Dong, Z., Chen, X., Xu, H., Wen, J.: Explainable legal case matching via inverse optimal transport-based rationale extraction. In: SIGIR (2022)
- 41. Zhang, H., Dou, Z., Zhu, Y., Wen, J.R.: Contrastive learning for legal judgment prediction. ACM Trans. Inf. Syst. **41**(4), 25 (2023)
- Zhao, Z., Chen, H., Zhang, J., Zhao, X., Liu, T., Lu, W., Chen, X., Deng, H., Ju, Q., Du, X.: Uer: An open-source toolkit for pre-training models. EMNLP-IJCNLP (2019)

43. Zhong, H., Wang, Y., Tu, C., Zhang, T., Liu, Z., Sun, M.: Iteratively questioning and answering for interpretable legal judgment prediction. In: AAAI (2020)