# Lawformer: A Pre-trained Language Model for Chinese Legal Long Documents

**Chaojun Xiao[1*], Xueyu Hu[2*], Zhiyuan Liu[1†], Cunchao Tu[3], Maosong Sun[1]**

[1]Department of Computer Science and Technology
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology, China
[2] Beihang University, Beijing, China
[3] Beijing Powerlaw Intelligent Technology Co., Ltd., China
xcjthu@gmail.com, huxueyu@buaa.edu.cn
tucunchao@gmail.com {liuzy,sms}@tsinghua.edu.cn

## Abstract

Legal artificial intelligence (LegalAI) aims to benefit legal systems with the technology of artificial intelligence, especially natural language processing (NLP). Recently, inspired by the success of pre-trained language models (PLMs) in the generic domain, many LegalAI researchers devote their effort to apply PLMs to legal tasks. However, utilizing PLMs to address legal tasks is still challenging, as the legal documents usually consist of thousands of tokens, which is far longer than the length that mainstream PLMs can process. In this paper, we release the Longformer-based pre-trained language model, named as Lawformer, for Chinese legal long documents understanding. We evaluate Lawformer on a variety of LegalAI tasks, including judgment prediction, similar case retrieval, legal reading comprehension, and legal question answering. The experimental results demonstrate that our model can achieve promising improvement on tasks with long documents as inputs. The code and parameters are available at https://github.com/thunlp/LegalPLMs.

## 1 Introduction

Legal artificial intelligence (LegalAI) focuses on applying methods of artificial intelligence to benefit legal tasks (Zhong et al., 2020a), which can help improve the work efficiency of legal practitioners and provide timely aid for those who are not familiar with legal knowledge. Thus, LegalAI has received great attention from both natural language processing (NLP) researchers and legal professionals (Zhong et al., 2018, 2020b; Wu et al., 2020; Chalkidis et al., 2020; Hendrycks et al., 2021).

In recent years, pre-trained language models (PLMs) (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020; Brown et al.,
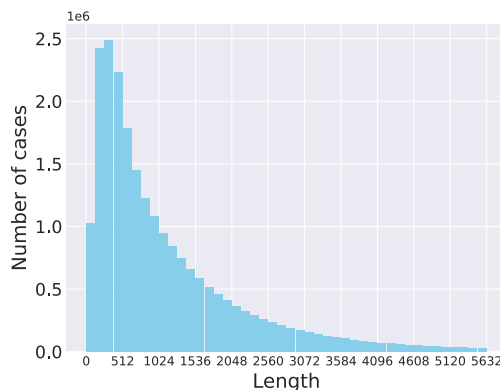


Figure 1: The length distribution of the fact description in criminal and civil cases. The criminal and civil cases consist of 1260.2 tokens on average. The length of 64.12% cases is over 512.

2020) have proven effective in capturing rich language knowledge from large-scale unlabelled corpora and achieved promising performance improvement on various downstream tasks. Inspired by the great success of PLMs in the generic domain, considerable efforts have been devoted to employing powerful PLMs to promote the development of LegalAI (Shaghaghian et al., 2020; Shao et al., 2020; Chalkidis et al., 2020).

Some researchers attempt to transfer the contextualized language models pre-trained on the generic domain, such as Wikipedia, Children's Books, to tasks in the legal domain (Shaghaghian et al., 2020; Zhong et al., 2020b; Shao et al., 2020; Elwany et al., 2019). Besides, some works conduct continued pre-training on legal documents to bridge the gap between the generic domain and the legal domain (Zhong et al., 2019; Chalkidis et al., 2020).

However, most of these works adopt the full self-attention mechanism to encode the documents and cannot process long documents due to the high computational complexity. As shown in Figure 1, the average length of the criminal cases and

---

[*]Indicates equal contribution.
[†]Corresponding author.

civil cases is 1260.2, which is far longer than the maximum length that mainstream PLMs (BERT, RoBERTa, etc.) can handle. With the limited capacity to process long sequences, these PLMs cannot achieve satisfactory performance in representing the legal documents (Zhong et al., 2020a; Shao et al., 2020). Hence, how to utilize PLMs to process legal long sequences needs more exploration.

In this work, we release Lawformer, which is pre-trained on large-scale Chinese legal long case documents. Lawformer is a Longformer-based (Beltagy et al., 2020) language model, which can encode documents with thousands of tokens. Instead of employing the standard full self-attention, we combine the local sliding window attention and the global task motivated full attention to capture the long-distance dependency. To the best of our knowledge, Lawformer is the first legal pre-trained language model, which can process the legal documents with thousands of tokens.

Besides, we evaluate Lawformer on a collection of typical legal tasks including legal judgment prediction, similar case retrieval, legal reading comprehension, and legal question answering. Solving these tasks requires the model to understand domain knowledge and concepts in legal texts, and be able to analyze complicated case scenarios and legal provisions.

Notably, the data distribution of existing datasets for legal judgment prediction is quite different from the real-world data distribution. And these datasets only contain criminal cases and omit civil cases. Therefore, we construct new datasets from scratch for legal judgment prediction tasks, which consist of hundreds of thousands of criminal cases and civil cases. As for the other tasks, we rely on the pre-existing datasets. Experimental results on these various tasks demonstrate that the proposed Lawformer can achieve strong performance in legal documents understanding.

The main contributions of this paper are summarized as follows:

- We release a Chinese legal pre-trained language model, which can process documents with thousands of tokens, named as Lawformer. To the best of our knowledge, Lawformer is the first pre-trained language model for legal long documents.

- We evaluate Lawformer for legal documents understanding, with high coverage of existing typical LegalAI tasks. In this benchmark, we propose

new legal judgment prediction datasets for both criminal and civil cases.

- Extensive experiments demonstrate the proposed Lawformer can achieve strong performance on various LegalAI tasks that require the models are able to process the long documents. In terms of the tasks with short inputs, Lawformer can also achieve comparable results with RoBERTa (Liu et al., 2019) pre-trained on the legal corpora.

## 2 Related Work

### 2.1 Legal Artificial Intelligence

Legal artificial intelligence aims to profit the tasks in the legal domain (Zhong et al., 2020a). Due to the amount of textual legal resources, LegalAI has drawn great attention from NLP researchers in recent years (Luo et al., 2017; Ye et al., 2018; Duan et al., 2019b; Shao et al., 2020; Wu et al., 2020). Early works attempt to analyze legal documents with hand-crafted features and statistical methods (Kort, 1957; Nagel, 1963; Segal, 1984). With the development of deep learning, many efforts have been devoted to solving various legal tasks, such as legal charge prediction (Luo et al., 2017; Zhong et al., 2018; Xiao et al., 2018; Chalkidis et al., 2019; Yang et al., 2019), relevant law article retrieval (Chen et al., 2013; Raghav et al., 2016), court view generation (Ye et al., 2018; Wu et al., 2020), reading comprehension (Duan et al., 2019a), question answering (Zhong et al., 2020b; Kien et al., 2020), and case retrieval (Raghav et al., 2016; Shao et al., 2020). Besides, inspired by the great success of PLMs in the generic domain, there are also some researchers conducting pre-training on legal corpora (Zhong et al., 2019; Chalkidis et al., 2020). However, these models adopt the BERT as the basic encoder, which cannot process documents longer than 512. To the best of our knowledge, Lawformer is the first pre-trained language model for legal long documents.

### 2.2 Pre-trained Language Model

Pre-trained language models (PLMs), which are trained on amounts of unlabelled corpora, are able to benefit a variety of downstream NLP tasks (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020). Then we will introduce previous works related to ours from the following two aspects: domain-adaptive pre-training and long-document pre-training.
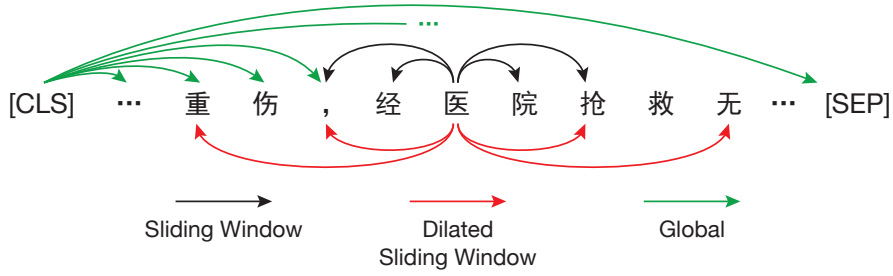
Figure 2: An example of the combination of the three types of attention mechanism in Lawformer. The size of the sliding window attention is 2. The size of the dilated sliding window attention is 2 and the gap is 1. The token, `[CLS]`, is selected to perform the global attention.

**Domain-Adaptive Pre-training.** Many researchers attempt to achieve performance gain by domain-adaptive pre-training on various domains, including the biomedical domain (Lee et al., 2020), the clinical domain (Alsentzer et al., 2019), the scientific domain (Beltagy et al., 2019), and the legal domain (Zhong et al., 2019; Chalkidis et al., 2020). These works further pre-train the BERT (Devlin et al., 2019) on the specific domain texts and have shown that continued pre-training on the target domain corpora can consistently achieve performance improvement (Gururangan et al., 2020).

**Long-Document Pre-training.** Due to the high computational complexity of the full self-attention mechanism, the traditional transformer-based PLMs are limited in processing long documents (Beltagy et al., 2020). Some works propose to utilize the left-to-right auto-regressive objective to pre-train the language models (Dai et al., 2019; Sukhbaatar et al., 2019). And some works attempt to reduce the computational complexity with sliding window based self-attention (Beltagy et al., 2020; Qiu et al., 2020; Zaheer et al., 2020). In this work, we adopt the widely-used Longformer (Beltagy et al., 2020) as our basic encoder, which combines the sliding window attention and the global full attention to process the documents.

## 3 Our Approach

### 3.1 Lawformer

Our current model utilizes Longformer (Beltagy et al., 2020) as our basic encoder. Instead of utilizing the full self-attention mechanism, Longformer combines the sliding window attention, dilated sliding window attention, and the global attention mechanism to encode the long sequence. We will introduce the three types of attention patterns in the section. Figure 2 gives an example of the combination of three types of attention mechanisms.

**Sliding Window Attention.** In this attention pattern, we only compute the attention scores between the surrounding tokens. Specifically, given the size of the sliding window $w$, each token only attends to the $\frac{1}{2}w$ tokens on each side. While the tokens will only aggregate information around them in each layer, with the number of layers increases, the global information can also be integrated into the hidden representations of each token.

**Dilated Sliding Window Attention.** Similar to dilated CNNs (van den Oord et al., 2016), the sliding window attention can be dilated to reach a longer context. In this attention mechanism, each window is not continuous but has a gap $d$ between each attended token. Notably, in the multi-head attention, the gap in different heads can be different, which can promote the model performance.

**Global Attention.** In some specific tasks, we need some tokens to attend to the whole sequence to obtain sufficient information. For example, in the text classification task, the special token `[CLS]` should be used to attend to the whole document. In the question answering task, the questions are supposed to attend the whole sequence to generate expressive representations. Therefore, we apply global attention for some pre-selected tokens for task-specific representations. That is, instead of attending to the surrounding tokens, the selected tokens will attend to the whole sequence to generate the hidden representations. Notably, the parameters in the global attention and sliding window attention are different.

With the three types of attention mechanisms, we can process the long sequences with linear complexity.

| Model | criminal | | | | | civil | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mic@c | Mac@c | Mic@l | Mac@l | Dis@t | Mic@c | Mac@c | Mic@l | Mac@l |
| BERT | 94.8 | 68.2 | 81.5 | 52.9 | 1.286 | 80.6 | 47.6 | 61.7 | 31.6 |
| RoBERTa | 94.7 | 69.3 | 81.1 | 53.5 | 1.291 | 80.0 | 47.2 | 60.2 | 29.9 |
| L-RoBERTa | 94.9 | 70.8 | 81.1 | 53.4 | 1.280 | 80.8 | 49.4 | 61.2 | 31.3 |
| Lawformer | **95.4** | **72.1** | **82.0** | **54.3** | **1.264** | **81.1** | **50.0** | **63.0** | **33.0** |

Table 1: The results on the legal judgment prediction dataset, CAIL-Long. For criminal cases, we evaluate the models on charge prediction task (Mic@c, Mac@c), relevant law prediction task (Mic@l, Mac@l), and term of penalty prediction task (Dis@t). For civil cases, we evalute the models on cause of actions prediction task (Mic@c, Mac@c), and relevant law prediction task (Mic@l, Mac@l).

## 3.2 Data Processing

We collect tens of millions of case documents published by the Chinese government from China judgment Online[1]. As the downstream tasks are mainly in the areas of criminal and civil cases, we only keep the documents of criminal cases and civil cases. We divide each document into four parts: the information about the parties, the fact description, the court views, and the judgment results. We only keep the documents with the fact description longer than 50 tokens. After the data processing, the rest of the data are used for pre-training. The detailed statistics is listed in Table 2.

| | # Doc. | Len. | Size |
|---|---|---|---|
| criminal cases | 5,428,717 | 962.84 | 17 G |
| civil cases | 17,387,874 | 1,353.03 | 67 G |

Table 2: The statistics of the pre-training data. # Doc. refers to the number of documents. Len. refers to the average length of the documents. Size refers to the size of the pre-training data.

## 3.3 Pre-training Details

Following the previous work (Beltagy et al., 2020), we pre-train Lawformer with MLM objective, continuing from the checkpoint, RoBERTa-wwm-ext, released in Cui et al. (2019). We set the learning rate as $5 \times 10^{-5}$, the sequence length as $4,096$, and the batch size as 32. As the length of legal documents is usually smaller than $4,096$, we concatenate different documents together to make full use of the input length. We pre-train Lawformer for $200,000$ steps, and the first $3,000$ steps are for warm-up. We utilize Adam (Kingma and Ba, 2015) to optimize the model. The rest of the hyper-parameters are the same as Longformer. We

pre-train Lawformer with $8 \times 32$G NVIDIA V100 GPUs.

In the fine-tuning stage, we select different tokens to conduct the global attention mechanism. For the classification task, we select the token, `[CLS]`, to perform the global attention. And for the reading comprehension task and the question answering task, we perform the global attention on the whole questions. For the specific details of each task, please refer to the next section.

## 4 Experiments

### 4.1 Baseline Models

To verify the effectiveness of the proposed model, we compare Lawformer with following competitive baseline models:

- BERT (Devlin et al., 2019): we simply fine-tune the published checkpoint, BERT-base-chinese, which is pre-trained on Chinese wikipedia documents[2], on the following downstream datasets.

- RoBERTa-wwm-ext (RoBERTa) (Cui et al., 2019): it is pre-trained with the whole word masking strategy, in which the tokens that belong to the same word will be masked simultaneously. Notably, Lawformer is pre-trained continuously from the RoBERTa.

- Legal RoBERTa (L-RoBERTa): we pre-train a RoBERTa (Liu et al., 2019) on the same legal corpus, continuing from the released RoBERTa-wwm-ext checkpoint.

As these baseline models can only process documents with less than 512 tokens, we only truncate the documents to 512 tokens for these models.

### 4.2 Legal judgment Prediction

**Dataset Construction:** Legal judgment prediction aims to predict the judgment results given

---

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

[2]https://zh.wikipedia.org/

| Model | P@5 | P@10 | P@20 | P@30 | NDCG@5 | NDCG@10 | NDCG@20 | NDCG@30 | MAP |
|---|---|---|---|---|---|---|---|---|---|
| BERT | 44.27 | 41.83 | 36.73 | 33.49 | 78.18 | 80.06 | 84.43 | 91.46 | 50.65 |
| RoBERTa | 45.93 | 41.71 | 36.53 | 33.40 | 79.93 | 80.57 | 84.99 | 91.82 | 50.77 |
| L-RoBERTa | 45.75 | 42.85 | 37.79 | 33.58 | 78.90 | 81.01 | 85.26 | 91.70 | 50.17 |
| Lawformer | **51.91** | **46.44** | **37.95** | **33.99** | **83.11** | **84.05** | **87.06** | **93.22** | **57.36** |

Table 3: The results on the legal case retrieval dataset, LeCaRD. The dataset contains long cases with thousands of tokens as candidates.

the fact description. It is a critical application in the LegalAI field and has received great attention recently. To facilitate the development of this task, Xiao et al. (2018) have proposed a large-scale criminal judgment prediction dataset, CAIL2018. However, the average case length of CAIL2018 is much shorter than that of real-world cases. Besides, CAIL2018 only focuses on criminal cases and omits civil cases. In this paper, we propose a new judgment prediction dataset, CAIL-Long, which contains both civil and criminal cases with the same length distribution as in the real world.

CAIL-Long consists of $1,129,053$ criminal cases and $1,099,605$ civil cases. For both criminal and civil cases, we take the fact description as inputs and extract the judgment annotations with regular expressions. Specifically, each criminal case is annotated with the charges, the relevant laws, and the term of penalty. Each civil case is annotated with the causes of actions and the relevant laws. The detailed statistics of the dataset are shown in Table 4.

| | # Case | Len. | # C. | # Law | prison |
|---|---|---|---|---|---|
| criminal | 115,849 | 916.57 | 201 | 244 | 0-180 |
| civil | 113,656 | 1,286.88 | 257 | 330 | – |

Table 4: The statistics of the judgment prediction datset, CAIL-Long. # Case denotes the number of cases. Len. denotes the average length of the fact description. # C. denotes the number of charges/cause of actions. # Law denotes the number of relevant laws. And prison denotes the range of term of penalty (unit: month).

**Implementation Details:** Following previous work (Zhong et al., 2018), we train the models in the multi-task paradigm. For criminal cases, the charge prediction and the relevant law prediction are formalized as multi-label text classification tasks. The term of penalty prediction task is formalized as the regression task. For civil cases, the cause of actions prediction is formalized as a single-label text classification task, and the relevant law prediction is formalized as a multi-label text clas-

sification task. The models for criminal and civil cases are trained separately.

We adopt micro-F1 scores (Mic@{c,l}) and macro-F1 scores (Mac@{c,l}) as metrics for classification tasks, and the log distance (Dis@t) as metric for the regression task. We set the learning rate as $5 \times 10^{-5}$, and batch size as 32.

**Result:** We present the results in Table 1. As shown in the table, Lawformer achieves the best performance among the four models in both the micro-F1 and macro-F1 scores. The improvement indicates that Lawformer can accurately capture the key information from the long fact description. Besides, the case numbers of different labels (charges, cause of actions, and laws) are highly imbalanced, and Lawformer can also achieve significant improvement in macro-F1 scores, which indicates that Lawformer is able to handle the labels with limited cases. However, the overall results are still unsatisfactory. It still needs more exploration to predict the judgment results more accurately and robustly.

### 4.3 Legal Case Retrieval

**Dataset:** Legal case retrieval is the specialized information retrieval task in the legal domain, which aims to retrieve similar cases given the query fact description. For this task, we adopt the Legal Case Retrieval Dataset (LeCaRD)[3] as our benchmark, which contains 107 query cases and $10,716$ candidate cases. Notably, the length of the candidate cases is extremely long. As shown in the Table 5, the average length of the cases is $6,319.14$, which is a great challenge for the existing models.

**Implementation Details:** We train the models with the binary classification task, which requires the models to judge whether the given candidate case is relevant to the query case. We set the fine-tuning batch size as 32, the learning rate as $10^{-5}$ for all models. For the baseline models, which can only process sequences with lengths no more than

---

[3]https://github.com/myx666/LeCaRD

| # Query | # Cand. | Q-Len. | C-Len. | # Pair |
|---------|---------|--------|--------|--------|
| 107 | 10,716 | 444.55 | 6,319.14 | 1,094 |

Table 5: The statistics of LeCaRD dataset. # Query and # Cand. denote the number of query cases and candidate cases. Q-Len. and C-Len. denote the average length of the fact description of the query cases and candidate cases. # Pair denotes the number of positive query-candidate case pairs.

512, we set the maximum length of the query and the candidates as 100 and 409, respectively. And for the Lawformer, we set the maximum length of the query and the candidates as 509 and 3, 072, respectively. All tokens of the query case are selected to perform the global attention mechanism.

Following the previous works, we adopt 5-fold cross-validation for the dataset. We employ the top-$k$ Normalized Discounted Cumulative Gain (NDCG@$k$), Precision (P@$k$), and Mean Average Precision (MAP) as evaluation metrics. For each model, we report the performance of the checkpoint with the highest average score on P@10, NDCG@10, and MAP.

**Result:** The results are shown in Table 3. We can observe that Lawformer can significantly outperform all baseline models. For example, Lawformer improves upon baselines by 6.59 points in terms of mean average precision. The similar case retrieval task requires the models to compare the query case and candidate case in detail. The baseline models even cannot read the complete documents, and thus perform unsatisfactorily. Lawformer can process sequences with thousands of tokens and achieve promising results. However, the average length of candidate cases in LeCaRD is 6, 319.14, which is also beyond the capacity of Lawformer. We argue that it still needs further exploration to retrieve the long cases.

### 4.4 Legal Reading Comprehension

**Dataset:** We utilize the Chinese judicial reading comprehension (CJRC) as our benchmark dataset for legal reading comprehension (Duan et al., 2019a). CJRC is published in the Chinese AI and Law Challenge contest. We adopt the dataset published in 2020 as the benchmark[4]. CJRC consists of 9, 532 question-answer pairs with corresponding supporting sentences. There are three types of answers for these questions, including the span of words, yes/no, and unanswerable. The detailed

---
[4] https://github.com/china-ai-law-challenge/CAIL2020

statistics of CJRC are shown in Table 6. It is worth mentioning that the average length of the documents in CJRC is only 441.04.

| # Doc | Len. | # Que. | # S-Que. | # YN-Que. | # U-Que. |
|-------|------|--------|----------|-----------|----------|
| 9,532 | 441.04 | 9,532 | 6,692 | 1,892 | 948 |

Table 6: The statistics of CJRC dataset. # Doc denotes the number of case documents. Len. denotes the average length of the documents. # Que., # S-Que., # YN-Que., and # U-Que. denotes the total number of questions, questions with span of words as answers, questions with yes/no answers, unanswerable questions, respectively.

**Implementation Details:** We implement the models following the source code provided in Duan et al. (2019a). We train the models to predict the start positions and end positions. And the supporting sentence prediction is formalized as the binary classification for all the sentences. For the Lawformer, we select the whole question to perform the global attention. We adopt the exact match score (EM) and F1 score as evaluation metrics.

**Result:** The results are shown in Table 7. From the results, we can observe that (1) L-RoBERTa and Lawformer, which are pre-trained on the in-domain corpus, can significantly outperform the BERT and RoBERTa model, which are pre-trained on the out-of-domain corpus. (2) L-RoBERTa can achieve comparable performance with Lawformer, as the long documents are filtered out when constructing the dataset. We argue that with the improvement of the ability to process long sequences, Lawformer can achieve better performance in the long document reading comprehension.

| Model | Answer | | Support | | Joint | |
|-------|--------|-------|---------|-------|-------|-------|
|       | EM | F1 | EM | F1 | EM | F1 |
| BERT | 50.70 | 66.53 | 31.45 | 71.80 | 21.19 | 52.15 |
| RoBERTa | 53.81 | 68.86 | 34.60 | 73.63 | 23.76 | 55.62 |
| L-RoBERTa | 54.12 | 69.76 | **35.73** | **74.44** | **24.42** | 56.40 |
| Lawformer | **55.02** | **69.98** | 35.15 | 74.28 | 24.18 | **56.62** |

Table 7: The results on the legal reading comprehension task. Answer refers to the question answering task. Support refers to the supporting sentences prediction task. Joint refers to the overall performance.

### 4.5 Legal Question Answering

**Dataset:** Legal question answering requires the model to understand legal knowledge and answer the given questions. A high-quality legal question answering system can provide an accurate le-

gal consult service. For this task, we select JEC-QA (Zhong et al., 2020b) to evaluate the performance of the proposed model. JEC-QA consists of $28,641$ multiple-choice questions from the Chinese national bar exam, which is quite challenging for existing models (Zhong et al., 2020b).

**Implementation Details:** We formalize the task as a text classification task. First, we concatenate the questions and candidate choices together to form the inputs of the models. Then a linear layer is applied to compute the matching scores of the candidates. Previous works also take the reading materials retrieved by statistical methods (BM25, TFIDF) as inputs (Zhong et al., 2020b,a). We ignore these reading materials in our experiments, as the retrieval results are unsatisfactory and would harm the performance. We set the learning rate as $2 \times 10^{-5}$ and the batch size as 32. The positions are set as 0 for questions, and 1 for candidate choices.

| | dev | | test | |
| Model | single | all | single | all |
| --- | --- | --- | --- | --- |
| BERT | 42.78 | 25.77 | 41.23 | 24.50 |
| RoBERTa | 44.46 | **28.18** | 43.18 | 27.09 |
| L-RoBERTa | 45.17 | 27.54 | 43.29 | **27.81** |
| Lawformer | **45.81** | 27.21 | **45.81** | 27.43 |

Table 8: The results on the legal question answering task. We adopt accuracy as the evaluation metric. Here, single denotes the single-answer questions and all denotes all questions.

**Result:** The results are shown in Table 8. As the inputs in the dataset do not require long-distance understanding, L-RoBERTa can achieve comparable results with Lawformer. However, as the task needs the models to perform complex reasoning (Zhong et al., 2020b), all the models cannot achieve promising results. Therefore, it remains a great challenge for future works to enhance the models with legal knowledge and the logic reasoning capacity.

## 5 Conclusion and Future Work

In this paper, we pre-train a Longformer-based language model with tens of millions of criminal and civil case documents, which is named as Lawformer. Then we evaluate Lawformer on several typical LegalAI tasks, including legal judgment prediction, similar case retrieval, legal reading comprehension, and legal question answering. The results demonstrate that Lawformer can achieve significant performance improvement on tasks with long

sequence inputs.

Though Lawformer can achieve performance improvement for legal documents understanding, the experimental results also show that the challenges still exist. In the future, we will further explore the legal knowledge augmented pre-training. It is an established fact that enhancing the models with legal knowledge is quite necessary for many LegalAI tasks (Zhong et al., 2020a).

Meanwhile, we will also explore the generative legal pre-trained model. In real-world legal practice, the practitioners need to conduct heavy and redundant paper writing works. A powerful generative legal pre-trained language model can be beneficial for legal professionals to improve work efficiency.

To summarize, we release Lawformer for legal long document understanding in this paper. In the future, we will attempt to integrate legal knowledge into the legal pre-trained language models, and pre-train generative models for LegalAI tasks.

## References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of Clinical Natural Language Processing Workshop*, pages 72–78.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of EMNLP-IJCNLP*, pages 3606–3611.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of ACL*, pages 4317–4323.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert:"preparing the muppets for court'". In *Proceedings of EMNLP: Findings*, pages 2898–2904.

Yen-Liang Chen, Yi-Hung Liu, and Wu-Liang Ho. 2013. A text mining approach to assist the general public in the retrieval of legal documents. *Journal of the American Society for Information Science and Technology*, 64(2):280–290.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for Chinese bert. *arXiv preprint arXiv:1906.08101*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of ACL*, pages 2978–2988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, et al. 2019a. Cjrc: A reliable human-annotated benchmark dataset for Chinese judicial reading comprehension. In *Proceedings of CCL*, pages 439–451. Springer.

Xinyu Duan, Yating Zhang, Lin Yuan, Xin Zhou, Xiaozhong Liu, Tianyi Wang, Ruocheng Wang, Qiong Zhang, Changlong Sun, and Fei Wu. 2019b. Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning. In *Proceedings of CIKM*, pages 1361–1370.

Emad Elwany, Dave Moore, and Gaurav Oberoi. 2019. Bert goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding. In *Proceedings of NeurIPS Workshop on Document Intelligence*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*, pages 8342–8360.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of COLING*, pages 988–998.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Fred Kort. 1957. Predicting supreme court decisions mathematically: A quantitative analysis of the" right to counsel" cases. *The American Political Science Review*, 51(1):1–12.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

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

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of EMNLP*, pages 2727–2736.

Stuart S Nagel. 1963. Applying correlation analysis to case prediction. *Tex. L. Rev.*, 42:1006.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *Proceedings of ISCA Speech Synthesis Workshop*, pages 125–125.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Jiezhong Qiu, Hao Ma, Omer Levy, Wen-tau Yih, Sinong Wang, and Jie Tang. 2020. Blockwise self-attention for long document understanding. In *Proceedings of EMNLP: Findings*, pages 2555–2565.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

K Raghav, P Krishna Reddy, and V Balakista Reddy. 2016. Analyzing the extraction of relevant legal judgments using paragraph-level and citation information. *AI4JCArtificial Intelligence for Justice*, page 30.

Jeffrey A Segal. 1984. Predicting supreme court cases probabilistically: The search and seizure cases, 1962-1981. *The American political science review*, pages 891–900.

Shohreh Shaghaghian, Luna Yue Feng, Borna Jafarpour, and Nicolai Pogrebnyakov. 2020. Customizing contextualized language models for legal document reviews. In *Proceedings of IEEE Big Data*, pages 2139–2148. IEEE.

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

Sainbayar Sukhbaatar, Édouard Grave, Piotr Bojanowski, and Armand Joulin. 2019. Adaptive attention span in transformers. In *Proceedings of ACL*, pages 331–335.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of EMNLP*, pages 763–780.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Wenmian Yang, Weijia Jia, XIaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of IJCAI*.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of NAACL-HLT*, pages 1854–1864.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Proceedings of NeurIPS*.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of EMNLP*, pages 3540–3549.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of ACL*, pages 5218–5230.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. Jecqa: A legal-domain question answering dataset. In *Proceedings of the AAAI*, volume 34, pages 9701–9708.

Haoxi Zhong, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. 2019. Open Chinese language pretrained model zoo. Technical report.