

PILOT: Legal Case Outcome Prediction with Case Law

Lang Cao¹, Zifeng Wang¹, Cao Xiao², Jimeng Sun¹

¹University of Illinois Urbana-Champaign

²GE Healthcare

{langcao2, zifengw2, jimeng}@illinois.edu, cao.xiao@ge.com

Abstract

Machine learning shows promise in predicting the outcome of legal cases, but most research has concentrated on civil law cases rather than case law systems. We identified two unique challenges in making legal case outcome predictions with case law. First, it is crucial to identify relevant precedent cases that serve as fundamental evidence for judges during decision-making. Second, it is necessary to consider the evolution of legal principles over time, as early cases may adhere to different legal contexts. In this paper, we proposed a new framework named PILOT (PredictIng Legal case OuTcome) for case outcome prediction. It comprises two modules for relevant case retrieval and temporal pattern handling, respectively. To benchmark the performance of existing legal case outcome prediction models, we curated a dataset from a large-scale case law database. We demonstrate the importance of accurately identifying precedent cases and mitigating the temporal shift when making predictions for case law, as our method shows a significant improvement over the prior methods that focus on civil law case outcome predictions.

1 Introduction

Predicting legal case outcomes is a crucial task that facilitates data-driven decision-making in legal cases based on relevant information, such as the factual description (Cui et al., 2022). With a significant number of legal cases arising worldwide each year, legal professionals face the daunting task of reviewing the extensive legal text and delivering accurate and fair outcomes. Legal case outcome prediction has the potential to simplify this labor-intensive document review process, enhancing strategy and decision-making. As the volume and complexity of cases continue to escalate, the development of precise and reliable legal case outcome prediction systems becomes an urgent priority.

Two legal frameworks exist across the globe: the *civil law system*, which assesses each case based on comprehensive codes and statutes, and the *case law system*, where the interpretation and application of law heavily depend on precedent court decisions. Most existing works were proposed for the civil law framework, including charge prediction, violated articles prediction, prison terms prediction, court decision prediction, and court view generation (Paul et al., 2020; Hu et al., 2018; Chen et al., 2019; Chalkidis et al., 2019a; Alali et al., 2021; Ye et al., 2018). However, predicting case outcomes in case law systems presents unique challenges distinct from those in civil law: (1) it requires the identification of similar historical cases, and (2) meanwhile accounting for the evolution of legal principles over time.

- **Precedent Cases** In the case law system, the application of precedents plays a crucial role. To achieve accurate prediction of case outcomes, it is vital to identify past cases that exhibit similar legal principles, factual contexts, and key arguments. Moreover, how to effectively utilize the retrieved cases in the prediction of new case outcomes still requires further exploration.
- **Temporal Shift** One aspect that has not received sufficient attention in previous research is the temporal evolution of legal principles. We argue that it is crucial to not only comprehend the historical context and development of legal precedents but also to effectively capture and represent the temporal shifts of laws in predictive modeling.

To fill the gap, we proposed a new model named PILOT (PredictIng Legal case OuTcome) for case outcome prediction, which consists of two functional modules:

- **Case Retrieval** We initially train the module in an unsupervised manner to obtain text embed-

dings for cases. These embeddings are then used to query and select the most relevant precedent cases, which serve as additional inputs to our main model.

- **Temporal Pattern Mining** A temporal decay term is introduced to ensure the model captures the more recent patterns and explicitly learns to adapt to the temporal pattern change.

To facilitate this line of research, we established a new dataset named *ECHR2023*, which was extracted from the European Court of Human Rights (ECHR) database¹ with focusing on precedent cases and temporal concept shift. We evaluated the proposed PILOT model against state-of-the-art models on *ECHR2023*. The experiment results show that PILOT substantially outperforms existing works in several metrics. The two modules in PILOT effectively improve the performance in different aspects.

In summary, the main contributions of this paper are as follows

- We highlight the issue of Temporal Pattern Shift in legal AI tasks. This problem is important but is usually ignored in most previous works.
- We propose a new method, PILOT, which can effectively handle Temporal Pattern Shift based on characteristics of the case law system.
- We contribute a new dataset, *ECHR2023*, for legal case outcome prediction. PILOT achieves state-of-the-art performance on *ECHR2023*.

2 Related Work

Legal Case Outcome Prediction on civil law framework has been well studied, mainly focusing on predicting whether the case description violates existing legislation. Machine learning technologies, including multi-task learning (Feng et al., 2022), few-shot learning (Hu et al., 2018; He et al., 2019) has been adopted. Model explanation has been another focused (Jiang et al., 2018; Zhong et al., 2020; Ge et al., 2021; Chen et al., 2019; Ye et al., 2018; Wu et al., 2020).

In contrast, for case law systems that heavily relies on judicial decisions of relevant precedent cases rather than solely on constitutional law when rendering final case outcomes, there are relatively few studies due to the lack of scarcity

of large-scale, high-quality, and structured labeled data. For instance, (Chalkidis et al., 2019a) utilize HIER-BERT to first encode individual facts and then employ two layers of transformers to encode all the facts within a given case. (Chalkidis et al., 2021) generate rationales through a text encoder sub-network that reads the text, a rationale extraction sub-network that identifies the most important words via a binary mask, and a prediction sub-network that classifies a hard-masked version of the text. They also incorporate rationale constraints as regularizers. (Paul et al., 2020) employ a fact encoding layer to encode facts and a charge encoding layer to encode charges. Subsequently, they use a Matching Layer, which incorporates an attention mechanism, to predict the final charges for each case. (Malik et al., 2021) utilize a Hierarchical XLNet architecture to predict case outcomes and related interpretations. These efforts primarily focus on the classification of fundamental case outcomes. To the best of our knowledge, most of the existing works do not handle temporal pattern shift.

Temporal Pattern Shift arises due to changes in label distribution, meaning, and etc. Existing research approaches this issue from different angles. For example, (Zhao et al., 2022) analyzes the impact of temporal pattern shift on model explanations. (Sun et al., 2018) explored drift adaptation through transfer-based ensemble learning. Fan et al. (Fan et al., 2023) proposed to use two CONET networks to model the normalized parameters of historical and future windows separately, enabling normalization and prediction of future sequences. (Lu et al., 2023) introduced an out-of-domain representation learning approach utilizing adversarial learning to capture domain-specific segments and a domain-independent commonality representation. (Rosin and Radinsky, 2022) introduced Temporal Attention and trained a transformer-based model with additional time-based inputs. In the legal field, (Chalkidis and Søgaard, 2022) tackled temporal pattern shift in legal text classification by proposing Label-Wise Distributional Robust Optimization. This algorithm addresses temporal pattern shift stemming from class imbalance problems and enhances model robustness. However, the existing works are too general and are designed for simple scenes, so they do not perform well in adapting the more complex shift. There is still a lack of a comprehensive solution for legal models to adapt

¹<https://hudoc.echr.coe.int>

the shift in the legal field directly and naturally.

3 ECHR2023 Dataset

We build a novel dataset called *ECHR2023* that takes the special challenges in legal case outcome prediction with case law. This dataset is derived from the most recent ECHR database. The primary focus of *ECHR2023* is to investigate the issue of temporal pattern shifts in the legal domain.

Data Acquisition and Processing The data extracted from the ECHR database is of low quality and contains a substantial amount of noise. The case documents are often excessively long, surpassing 2,000 words, and may consist of text in multiple European languages. As a result, the readability and quality of the text data are poor, posing difficulties for humans in comprehending the content of the cases.

Specifically, we prompt the large language model, *gpt-3.5-turbo*², to process the raw data. The prompts guide the model to focus on the primary arguments in the case and summarize them into more concise points. Therefore, the output of LLMs will not introduce new information or fabrication to a case but rather retain the important parts of the original information. We employ LLMs to summarize original legal documents with the aim of simplifying the input and concentrating on identifying temporal pattern shifts. The prompts, example input-output of the LLM, and more details in the raw data processing can be found in Appendix B. The resulting sample is described by the following attributes: case ID, title of the case, outcome decision date of the case, corresponding legal article, and text description of the case. Following the processing results, we conduct a manual review of the generated summaries to ensure their quality and to eliminate any data that is obviously incorrect.

Data Analysis Most of existing datasets are random split and ignore the temporal pattern change in the real world. We analyze the temporal pattern change in this dataset as follows: we perform outcome prediction using BERT (Devlin et al., 2019) using both the random split and chronological split. For random split (that does not consider case time), the performance of model training in Micro-F1 is 0.798, and the testing performance is 0.796. While for chronological split data split that we train the model using previous cases and test on cases that happen later, the performance of model training

in Micro-F1 is 0.737 and the testing performance is 0.677, which shows the patterns learned from previous data cannot fully capture the signal in new cases.

4 Methodology

4.1 PILOT Framework

Problem 1 (Legal Case Outcome Prediction with Case Law). *Given a set of n chronological ordered cases $\mathbf{C} = \{C_i\}_i^n$, where each C_i is represented by the text description of the case, the legal case outcome prediction aims at predicting whether a new case violates any legal article in $\mathbf{V} = \{V_1, V_2, \dots, V_n\}$. Here $V_i \in \{0, 1\}^L$ is the corresponding multi-hot label vector of the case C_i violated articles, and L is the total number of law articles. This task is a multi-label classification to decide the case C_i violated law articles V_j .*

We propose PILOT that primarily focuses on two distinct challenges in predicting legal case outcomes with case law: effectively identifying similar precedent cases and handling temporal pattern shift of legal principles. As illustrated by Figure 1, PILOT consists of three modules: the Relevant Case Retrieval module that retrieves relevant cases as references for outcome prediction, the Case Encoder with Evidence Fusion module that uses encodes current case with fact description and relevant cases, and the Temporal Shift Mining module that is directly adapting to the temporal drift. We will now provide a detailed introduction to each of these modules.

4.2 Precedent Case Retrieval

In the case law system, precedent cases serve as crucial references that judges rely on when making decisions for new cases. In order to emulate this decision-making process, we develop a precedent case retrieval module that enhances case outcome prediction in two key aspects: (1) by providing augmented evidence for prediction and (2) by offering interpretability through the provision of evidence.

Case Encoding We execute contrastive learning based on a pre-trained language model on case documents only from training split of dataset. We suppose that we only have case documents in training split in the database at the beginning. The yielded model is then utilized for encoding all legal case documents in the database and the current case, preparing for similarity search. Formally, the contrastive learning is performed based on InfoNCE

²<https://platform.openai.com/docs/models/gpt-3-5>

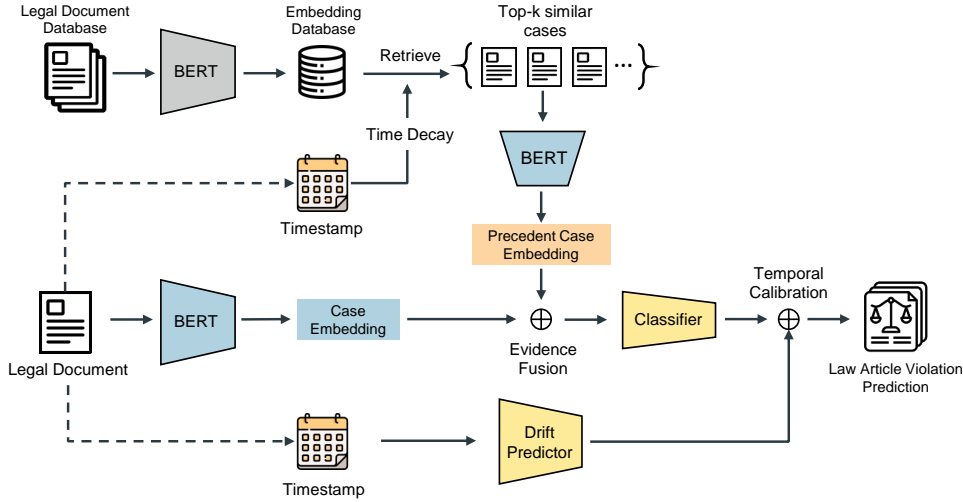


Figure 1: The framework of our proposed model PILOT. PILOT has three modules: Relevant Case Retrieval, Case Encoder with Evidence Fusion, and Temporal Shift Mining. The Relevant Case Retrieval module retrieves relevant cases to use as references for outcome prediction. The Case Encoder with Evidence Fusion module encodes current cases with fact descriptions and relevant cases. The Temporal Shift Mining module adapts directly to temporal drift.

loss (Gao et al., 2021). Without the need for annotated labels, each case is passed into the BERT model twice within a batch, resulting in two different document embeddings H_i^0 and H_i^1 due to the randomness of the dropout layers (Srivastava et al., 2014).

Within a batch, each pair (H_i^0, H_i^1) is positive, and all the other pairs that (H_i^0, H_j^1) where $i \neq j$ are negative. The contrastive training objective ℓ_i is hence defined by:

$$\ell_i = -\log \frac{e^{\cos(H_i^0, H_i^1)/\tau}}{\sum_{j=1}^N e^{\cos(H_i^0, H_j^1)/\tau}}, \quad (1)$$

where H_i is the document embedding of case C_i , H_i^0 is the positive sample for H_i , τ is a temperature hyperparameter, and $\cos(H_i, H_j)$ measures the cosine similarity of the input embeddings.

Case Retrieval After training with Eq. (1), we encode all legal cases in the database into semantically meaningful document embeddings $\mathbf{H} = \{H_1, H_2, \dots, H_n\}$, which can be used to compute cosine similarities for case retrieval.

In this work, we put temporal constraints for the retrieval process. First, the retrieval is performed considering the timestamps of the target cases because a case cannot refer to any future cases. For a case $C_i \in \mathcal{C}$, we assign the similarity $\text{sim}(C_i, C_j) = -1$ if $i < j$ to filter out future cases from the candidate pool.

Secondly, we also take into consideration the influence of temporal pattern shifts of legal principles,

as recent cases often carry higher reference value in legal decision-making. Based on this insight, we design a variant of cosine similarity equipped with a temporal decayed function as

$$\text{sim}(C_i, C_j) = \frac{\cos(C_i, C_j)}{1 + \frac{T_i - T_j}{\alpha \times |\mathcal{C}_{\text{val}}|}}, \quad (2)$$

where C_j is a candidate similar case, α is temporal decayed coefficient, and $|\mathcal{C}_{\text{val}}|$ is the size of validation split in the dataset. We set the decaying unit as the size of the validation split because it is a time span from labeled data to the newest unlabeled data, which is also the length of validation data. When $\alpha = (T_i - T_j)/|\mathcal{C}_{\text{val}}|$, the similarity score of (C_i, C_j) will be half. As α decreases, the reference value of precedent cases will decrease faster.

4.3 Case Encoding with Evidence Fusion

Target Case Encoding To prepare legal case data for outcome prediction, the first step is to embed the case documents into contextualized representation. To achieve this, we preprocess the legal document text data as follows: we convert the fact list to a piece of text by replacing all carriage return characters in the text with spaces, then use *BertTokenizer* to conduct tokenization.

Next, the preprocessed legal document text data is passed into a pre-trained language model (PLM) for further processing. Here we choose *legal-bert-base-uncased* (Chalkidis et al., 2020b), which is

pre-trained on different kinds of legal documents, enabling it to capture and understand the context and meaning of the text. For every case C_i , we pass it into the PLM and get the contextualized representation of the fact description $H_i \in \mathbb{R}^{d_t}$, where d_t is the dimension of the last hidden layer in PLM. We indicate this contextualized representation H_i as the current case embedding E_i :

$$E_i = H_i = PLM(C_i), \quad (3)$$

The PLM takes the preprocessed legal document text data as input and generates a contextualized representation of the legal case text, encapsulating the semantic and syntactic information of the legal fact description. It captures the relationships between words, phrases, and sentences, providing a rich representation of the text’s meaning within the legal context.

Evidence Fusion We use the target case C_i to query all cases \mathbf{C} to retrieve the top k similar precedent cases according to similarity scores computed by Eq. (2). We draw the evidence $\mathbf{R}_i = \{R_1, R_2, \dots, R_k\}$ from the retrieved cases, where $R_j = \{\text{sim}(C_i, C_j), V_j\}$ includes the case result $V_j \in \{0, 1\}^L$ and the similarity score $\text{sim}(C_i, C_j)$ of this relevant case.

Based on the evidence \mathbf{R}_i retrieved from precedent cases, we build the evidence embedding E_i^r by:

$$E_i^r = \sum_{j=1}^k \frac{e^{\text{sim}(C_i, C_j)} \times V_j}{\sum_{j=1}^k e^{\text{sim}(C_i, C_j)}}. \quad (4)$$

where $E_i^r \in \mathbb{R}^L$. We concatenate current case embedding E_i^c with relevant case embedding E_i^r to get the input of the linear classifier layer for C_i by:

$$E_i = [E_i^c, E_i^r], \quad (5)$$

where $E_i \in \mathbb{R}^{d_t+L}$.

This approach allows the model to learn the relationship between relevant cases, leading to a better understanding of the factors influencing case outcomes. Moreover, it helps alleviate the impact of temporal pattern shift by providing a local perspective that captures the evolving nature of legal precedents.

4.4 Outcome Prediction with Temporal Pattern Mining

To further mitigate the temporal pattern drift when the model makes outcome predictions, we introduce a drift prediction module that mines the effect

of timestamps to the final outcomes:

$$\text{Drift}_i = \text{MLP}(T_i), \quad (6)$$

where $\text{Drift}_i \in \mathbb{R}^d$. MLP is a two-layer multi-layer perceptron, and the dimension of the hidden layer is d . We add the output Drift_i to the original prediction to get the final prediction:

$$y_i^{\text{final}} = y_i^{\text{orig}} + \text{Drift}_i, \quad (7)$$

where y_i^{orig} is original output generated by the classifier and $y_i^{\text{final}} \in \mathbb{R}^L$. The drift prediction module explicitly incorporates a global view by adapting to the temporal concept and learning from the entire timeline. By considering the evolution of legal precedents over time, this module effectively captures and adapts to the changes in the legal landscape, ensuring that the model remains robust and accurate in predicting case outcomes.

4.5 Training and Loss Function

In addition to the binary cross-entropy loss \mathcal{L}_{BCE} used for the multi-label classification task, we add the drift loss $\mathcal{L}_{\text{Drift}}$ to the model loss function. $\mathcal{L}_{\text{Drift}}$ uses mean squared error loss to calculate the drift distance between original predictions and final predictions. The loss function of this model is defined as:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{BCE} + \lambda \sum_{i=1}^L (y_i^{\text{final}} - y_i^{\text{orig}})^2, \quad (8)$$

where λ is the weight that balances the two losses.

5 Experiments

In this section, we conducted extensive experiments to show the performance of PILOT associated with more in-depth analysis. Universally, we report the average results of all models obtained by five runs with different random seeds, to ensure fair comparison. We use four metrics to evaluate the legal case outcomes: micro-F1, micro-Jaccard, micro-PR-AUC, and micro-ROC-AUC. More training details can be found in Appendix A.

As for the availability of cases during training and evaluation, we strictly ensure that we do not use any later cases as references for the current case. During the training phase, all prior cases from the training set are available as precedents. At test time, all prior cases from both the training and test sets are available. In the contrastive learning of the case encoding model, we only use data from the training split of the dataset.

Method	F1	Jaccard	PR-AUC	ROC-AUC
BERT	0.675±0.005	0.509±0.005	0.498±0.004	0.795±0.011
HIER-BERT	0.680±0.008	0.516±0.009	0.502±0.011	0.803±0.004
BERT-LWAN	0.655±0.012	0.488±0.014	0.477±0.009	0.782±0.017
EPM-base	0.657±0.012	0.490±0.013	0.482±0.014	0.781±0.006
BERT+CL+kNN	0.679±0.006	0.514±0.007	0.502±0.006	0.793±0.015
BERT+TemporalAttention	0.648±0.009	0.480±0.010	0.459±0.012	0.791±0.008
LWDROV2	0.694±0.013	0.531±0.015	0.511±0.016	0.830±0.011
ChatGPT 5-shots	0.442	0.284	0.267	0.818
PILOT (Ours)	0.715±0.008	0.557±0.010	0.543±0.014	0.831±0.007

Table 1: Experimental results. The best results are in bold. PILOT significantly outperforms all other methods in all metrics. \pm represents standard deviation from five results of five different seeds.

5.1 Baselines

We consider the following baselines in evaluation.

- **BERT** (Devlin et al., 2019) is a transformer-based (Vaswani et al., 2017) language model pre-trained on large-scale web texts. We fine-tune and predict with the $[CLS]$ token of BERT.
- **HIER-BERT** (Chalkidis et al., 2019b) is a hierarchical version of BERT. This model was proposed to predict legal judgment for long documents by first splitting and encoding raw law documents into multiple sentence embeddings, then fusing them with a two-layer Transformer model (Vaswani et al., 2017) to yield the document embeddings.
- **BERT-LWAN** (Chalkidis et al., 2020a) is Label-Wise Attention Network after BERT that was shown to be robust in multi-label classification. LWAN employs L attention for L labels to learn the semantics of label interpretation.
- **EPM-base** (Feng et al., 2022) is the variant of the state-of-the-art method on the CAIL2018 dataset. The original model, named Event-based Prediction Model (EPM) targets Chinese legal case outcome prediction, augmented by extra annotations about the legal event information. We remove the event extraction module in our experiments for fair comparison and refer the method to the name EPM-base.
- **BERT+CL+kNN** (Su et al., 2022) is an advanced method for general purpose multi-label prediction. It is equipped with a k-nearest-neighbor model along with a multi-label contrastive learning objective for better multi-label classification performance.
- **BERT+TemporalAttention** (Rosin and Radinsky, 2022) adds a time-aware self-attention mod-

ule to the transformer model, which demonstrates superior performance in capturing temporal patterns when making predictions. In detail, it adds a time matrix to the attention weight to learn the impact of the temporal shift.

- **LWDROV2** (Chalkidis and Sogaard, 2022) was proposed for legal text classification tasks. It employs Label-Wise Distributional Robust Optimization to mitigate class imbalance and temporal pattern shift problems.
- **ChatGPT 5-shots** (Ouyang et al., 2022) is based on the in-context learning capability of the GPT-3.5-turbo model. To be specific, we put the exemplar cases and their outcomes retrieved using our Precedent Case Retrieval module into the context, then prompt the language model to generate the outcome predictions.

Evaluation Strategy. To ensure that future information is not used in legal case outcome predictions, we partitioned the data chronologically. As a result, the training, validation, and test data consist of 8,138, 3,000, and 3,000 instances, respectively, ensuring a preserved time span between the sets. In addition, this chronological split enables the evaluation of models’ adaptability to concept drift and reinforces temporal coherence. The dataset provides a substantial amount of validation and test data, contributing to its superior evaluation capabilities for legal case outcome prediction compared to existing alternatives. The statistics of the case outcomes are summarized in Table 2.

5.2 Result: Legal Case Outcome Prediction

We report the main results of legal case outcome predictions in Table 1. From the table, we observe that our method outperforms other methods by a

ECHR Articles	Train	Dev.	Test
Right to life	432	180	188
Prohibition of torture	1,048	796	835
Right to liberty and security	1,264	608	690
Right to a fair trial	4,969	1,165	1,081
No punishment without law	32	7	9
Right for private and family life	682	287	421
Freedom of religion	43	17	26
Freedom of expression	313	151	194
Freedom of assembly	104	80	148
Right to an effective remedy	1,202	506	520
Prohibition of discrimination	170	48	61
Derogation in time of emergency	4	9	10
Individual applications	58	46	60
Examination of the case	34	4	7
Protection of property	1,483	435	347
Signature and ratification	5	11	21

Table 2: Label distribution of the ECHR2023 dataset.

Method	F1	▽
PILOT	0.712	-
w/o relevant case retrieval	0.701	-0.011
w/o temporal pattern handling	0.697	-0.015
w/ law article semantics	0.705	-0.007

Table 3: Results of ablation study. Relevant case retrieval and temporal pattern handling bring improvement to the model respectively, while incorporating law articles semantics has a performance drop. ▽ means the performance drop comparing with the method PILOT.

large margin in four metrics, especially over the methods that do not explicitly consider the temporal pattern shifts in legal case outcomes.

In addition, our method improves the micro-F1 by 2.74% than the previous state-of-the-art method of legal outcome prediction, LWDROV2. The reason is that LWDROV2 is a general label-wise robust method that does not solve temporal shifts directly. By contrast, our method employs a time-aware drift prediction module and augments the predictions with precedent cases.

It is noteworthy that ChatGPT 5-shots exhibits lower performance when compared to other prediction models based on supervised learning. In many instances, ChatGPT refuses to provide predictions, leading to limitations in its ability to make accurate determinations. Consequently, there remains the potential for further advancements in general-purpose generative large language models for predicting legal outcomes.

5.3 Result: Ablation Study

We performed an ablation study to evaluate the impact of the relevant case retrieval module and the temporal pattern handling module on the overall performance of our model. Table 3 presents the results of this study, highlighting how these two modules contribute to the improvement of the base model in distinct ways.

Additionally, we explored the incorporation of law article semantics into the model, using techniques such as law side attention or similar approaches employed in previous methods. Surprisingly, our findings indicated a decrease in performance when integrating law article information into our model. This observation is supported by the results in Table 1, where both the EPM-base and BERT-LWAN models, which incorporate law article information, exhibited inferior performance compared to BERT alone. We think one reason incorporating law articles undermines the performance is that the content and interpretations of law articles change as time goes on. It will influence model prediction without considering the time factor.

5.4 Result: Qualitative Case Study for Case Retrieval

The relevant case retrieval module is utilized for retrieving the top k precedent cases that are relevant to the target case. In Table 4, we present an example of the retrieval results. It is evident from the table that these retrieved cases exhibit semantic relevance to the target case. Furthermore, the violated articles mentioned in the retrieved cases are closely related and encompass the violated articles of the target case, indicating a comprehensive coverage of relevant legal provisions. Therefore, it demonstrates the effect of the case retrieval process from a qualitative perspective.

5.5 Result: Hyperparameter Analysis for Case Retrieval Module

The relevant case retrieval module encompasses two hyperparameters. The first parameter, denoted as k , determines the number of top relevant precedent cases to be retrieved. The second parameter is the coefficient α associated with the temporal decayed function in Eq. (2). The experimental results, presented in Table 2, shed light on the impact of these hyperparameters.

From the results, we conclude that including

Case	case id	main text (selected sentences)	violated articles	similarity
Current Case	001-199268	the applicant complained about the lack of effective remedy in domestic law	["13", "6"]	-
Precedent Case 1	001-195868	the applicant expressed concerns about the lack of effective remedies in domestic law	["3", "13"]	0.597
Precedent Case 2	001-189950	applicant complained about inadequate detention conditions	["3", "13"]	0.560
Precedent Case 2	001-198818	applicant complained about the excessive length of civil proceedings	["13", "6"]	0.421
Precedent Case 3	001-199269	complaint concerns the length of administrative proceedings regarding social benefits	["6"]	0.380
Precedent Case 4	001-198820	the applicant complained about the excessive length of his pre-trial detention	["6", "5"]	0.364

Table 4: An example of similar case retrieval results.

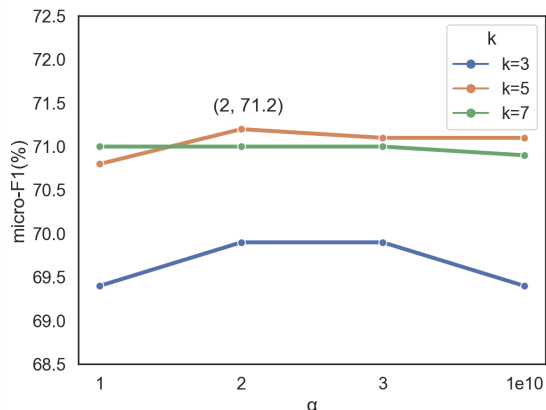


Figure 2: Hyperparameter analysis of k and α in the relevant case retrieval module. When k equals 5 and α equals 2, the model achieves the best results. When the value of α is $1e10$, it indicates an extreme condition that implies the absence of temporal decay in the computation of the similarity score

only three reference cases can introduce noise and lead to a decrease in performance, as it fails to retrieve the correct relevant cases effectively. However, utilizing five or seven reference cases demonstrates improved robustness compared to three cases. Notably, setting the value of α to $1e10$ is an extreme condition that implies the absence of temporal decay in the computation of the similarity score. The results indicate that incorporating the time-decayed function brings about some improvement over the original approach. Empirically, we find setting $\alpha \in [1, 10]$ yields the optimal results.

5.6 Result: Hyperparameter Analysis for Training Objective

To assess the impact of varying drift loss weights (λ), we conducted evaluations using different values. The results are presented in Figure 3. It is evident from the table that the inclusion of the drift loss contributes to improved model training and overall performance. Notably, the best value for λ , which balances the weighting between \mathcal{L}_{BCE} and \mathcal{L}_{Drift} , is found to be 0.10. The λ value of 0 indicates the exclusion of the drift loss from the model.

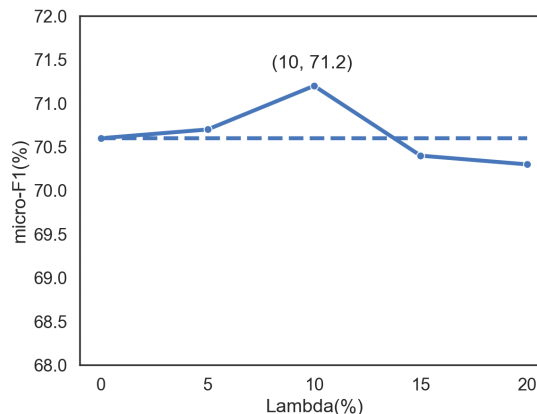


Figure 3: Hyperparameter analysis of lambda which is the weight of drift loss. When λ equals 0.10, the model achieves the best results.

Conversely, assigning a large value to λ can have a detrimental effect on the model's performance.

6 Conclusion

In conclusion, this paper introduces the PILOT model to tackle the challenges associated with predicting case outcomes in case law systems. Through our experiments, we have demonstrated the superior accuracy of our model in predicting case outcomes compared to existing methods. This improvement can be attributed to the identification of similar cases and the effective handling of temporal pattern changes.

Moreover, our proposed model goes beyond enhancing the accuracy of legal case predictions. It also offers valuable insights into legal reasoning and the evolution of legal principles. Precedent cases hold significant importance within the case law legal framework. It is worth noting that many previous works have primarily focused on the civil law system, which differs from the case law system. By analyzing and leveraging precedent cases, our model provides a deeper understanding of the underlying legal principles and their application.

Limitations

Deciding the outcome of legal cases is a very complex process in the real world. In this paper, we simplify many settings in real court scenarios to facilitate our research. The proposed model PILOT is a preliminary work in legal case outcome prediction, which might serve as a baseline for future investigation. The goal of designing the PILOT model is to highlight and alleviate the temporal pattern shift. There are many bias problems that need to be eliminated, and the model needs better interpretability to give reliable outcomes. It cannot be applied in the real world directly. Here are some ways to enhance the capability of PILOT before its application:

- More factors should be considered when designing a precedent case retrieval module. Currently, relevant cases are determined based on semantic similarity alone. However, relevant cases may not always be entirely semantically similar. Additionally, differences in factual details among cases can lead to different legal outcomes. Therefore, a more robust retrieval module with more retrieval factors should be developed if PILOT is to be applied in real-world scenarios.
- We need to further eliminate bias issues of PILOT before applied in real life.
- The model should prioritize better interpretability in order to provide reliable outcomes, given the need for transparency in the legal domain. For example, we can add a generation module let PILOT generate some explanation of its judgement.
- Legal outcomes should not be determined by a single model alone. Instead, a Mixture-Of-Experts approach can be employed, utilizing multiple instances of PILOT with varying hyperparameters, to perform ensemble learning and generate diverse results. After a voting process, the results can be more impartial.
- The model can benefit from incorporating more information from the case. Currently, only the factual section of the case is utilized, but additional information could be included to improve the model's performance.

Ethics Statement

Accuracy and Transparency. We are committed to ensuring the accuracy of our predictions to the best of our abilities. We will maintain transparency

about the methodologies, data sources, and algorithms used in our prediction models. We understand the profound implications of our work and strive to prevent any potential harm caused by inaccurate predictions.

Fairness and Impartiality. We pledge and strive to ensure our prediction models do not perpetuate or amplify any form of bias or discrimination. We will regularly audit our models to detect and mitigate any unfair bias, ensuring our predictions are objective and impartial.

Respect for Privacy and Confidentiality. We will strictly adhere to all applicable laws and regulations concerning data privacy and confidentiality. We will only use data that has been lawfully and ethically obtained, ensuring the privacy of all individuals involved is respected.

Accountability. We acknowledge our responsibility for the predictions made by our models. We will continually monitor and refine our models to ensure their reliability and validity.

Legal Compliance. We understand the significance of legal regulations and standards in our work. We will ensure full compliance with all relevant legal and professional guidelines in our legal outcome prediction task.

References

- Mohammad Alali, Shaayan Syed, Mohammed Alsayed, Smit Patel, and Hemanth Bodala. 2021. [JUSTICE: A benchmark dataset for supreme court's judgment prediction](#). *CoRR*, abs/2112.03414.
- Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019a. [Neural legal judgment prediction in English](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.
- Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019b. [Neural Legal Judgment Prediction in English](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.
- Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020a. [An Empirical Study on Large-Scale Multi-Label Text Classification Including Few and Zero-Shot Labels](#). In *Proceedings of the 2020*

- Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, Online. Association for Computational Linguistics.
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020b. [LEGAL-BERT: The muppets straight out of law school](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.
- Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. [Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.
- Ilias Chalkidis and Anders Søgaard. 2022. [Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454, Dublin, Ireland. Association for Computational Linguistics.
- Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. [Charge-based prison term prediction with deep gating network](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6362–6367, Hong Kong, China. Association for Computational Linguistics.
- Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. 2022. [A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges](#). ArXiv:2204.04859 [cs].
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. 2023. [Dish-ts: A general paradigm for alleviating distribution shift in time series forecasting](#). In *Proceedings of the AAAI conference on artificial intelligence*.
- Yi Feng, Chuanyi Li, and Vincent Ng. 2022. [Legal Judgment Prediction via Event Extraction with Constraints](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664, Dublin, Ireland. Association for Computational Linguistics.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jidong Ge, Yunyun Huang, Xiaoyu Shen, Chuanyi Li, and Wei Hu. 2021. [Learning fine-grained fact-article correspondence in legal cases](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3694–3706.
- Congqing He, Li Peng, Yuquan Le, Jiawei He, and Xiangyu Zhu. 2019. [Secaps: A sequence enhanced capsule model for charge prediction](#). In *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part IV*, page 227–239, Berlin, Heidelberg. Springer-Verlag.
- Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. [Few-shot charge prediction with discriminative legal attributes](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Xin Jiang, Hai Ye, Zhunchen Luo, WenHan Chao, and Wenjia Ma. 2018. [Interpretable rationale augmented charge prediction system](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 146–151, Santa Fe, New Mexico. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2017. [Fixing weight decay regularization in adam](#). *CoRR*, abs/1711.05101.
- Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, and Xing Xie. 2023. [Out-of-distribution representation learning for time series classification](#). In *The Eleventh International Conference on Learning Representations*.
- Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. [ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library.
- Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. 2020. **Automatic charge identification from facts: A few sentence-level charge annotations is all you need.** In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1011–1022, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Guy D. Rosin and Kira Radinsky. 2022. **Temporal Attention for Language Models.** In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting.
- Xi’ao Su, Ran Wang, and Xinyu Dai. 2022. **Contrastive Learning-Enhanced Nearest Neighbor Mechanism for Multi-Label Text Classification.** In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 672–679, Dublin, Ireland. Association for Computational Linguistics.
- Yu Sun, Ke Tang, Zexuan Zhu, and Xin Yao. 2018. Concept drift adaptation by exploiting historical knowledge. *IEEE transactions on neural networks and learning systems*, 29(10):4822–4832.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. **Transformers: State-of-the-Art Natural Language Processing.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- 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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780, Online. Association for Computational Linguistics.
- 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 the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.
- Zhixue Zhao, George Chrysostomou, Kalina Bontcheva, and Nikolaos Aletras. 2022. **On the impact of temporal concept drift on model explanations.** In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4039–4054, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. **Iteratively questioning and answering for interpretable legal judgment prediction.** *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1250–1257.

A Training Details

In the model training, we fine-tune on *legal-bert-base-uncased* (Chalkidis et al., 2020b). AdamW optimizer (Loshchilov and Hutter, 2017) was used to optimize the parameters of the model during the training. We apply differential learning rates. The learning rate of the final linear classifier is set to 1e-3, while others are all set to 1e-5. The Dropout (Srivastava et al., 2014) rate after the PLM output is set to 0.2. The batch size in each training step is set to 8. In training, we set an early stop strategy with 2 epochs. We use micro-F1 as monitoring indicators in our early stop strategy. We train *CaseSifter* with all data 3 epochs.

The code implementation of our model is mainly written using PyTorch (Paszke et al., 2019) library, and the pre-trained model is loaded using Transformers (Wolf et al., 2020) library. In addition, model training and evaluation were conducted on one NVIDIA GeForce RTX 3090.

B Raw Data Processing with LLMs

We use large language models (LLMs) to process raw data. The original document is lengthy and redundant. Our summarization target is the FACT section of case documents. We employ multiple regular expressions to filter out only the FACT section from the case documents and then input them into the LLM. We ensure that the input data does not contain any other parts of the case documents, which may leak information about the results. We prompt the *gpt-3-5-turbo* model to get output as processed data of a long document of one legal case. We utilized the default hyperparameters, setting the *temperature* to 1 and the *repetition_penalty* to 0. The maximum sequence length of the output is set to 512 tokens to ensure compatibility with BERT.

We have tried several prompts and select prompt according to summary performance of the model. The final selected prompt is shown in Figure 4. In our prompts, we guide the model to focus on the primary arguments in the case and summarize them into more concise points. Therefore, the output of LLMs will not introduce new information or fabrication to a case but rather retain the important parts of the original information. In this case, we can minimize the problem of hallucinations caused by generative language models as much as possible. We also acknowledge that this method will cause potential semantic loss in new dataset, but it can

increase the model inference speed and improve readability of original case documents.

We manually check the data quality from the LLM output. We review about dozens of samples of data. We observe that it do not introduce any new fabricated facts in the output, and indeed summarizes some key points of the case, which meets our expectations. We have also conducted experiments to compare these aspects. Our results show that using the baseline results of ChatGPT processed content only leads to a 0.5% decrease in performance than original lengthy documents, but significantly increases the training speed in later stages.

An example input and output of the LLM in data processing is shown in Figure 5.

Prompt

User: You are a good judge. This is a document about facts of a case. Help me find main descriptions of the facts in bullet points: {document of a legal case}

Figure 4: The final selected prompt. We also prompt model by telling *you are a good judge*.

Example

User: You are a good judge. This is a document about facts of a case. Help me find main descriptions of the facts in bullet points: A. Background of the case 4. The first applicant, Ms R.I., was born in 1975 and lives in Bucharest. She was in a stable relationship with R. She had two children with him: the second applicant, M.I., born in 2006, and the third applicant, I.I., born in 2009. In 2010, the relationship broke down and R. moved out of the family home.**(total 3618 tokens)**.....The break of the family ties between the mother and her children will without any doubt become permanent if the children are to live with their father, and therefore any visiting programme for the mother and the children will be illusory and formal in so far as the mother was unable to obtain the enforcement of a final decision for more than four years and as during that time she practically could not even speak with the children.”

ChatGPT: - Ms R.I. had two children with R.

- The relationship between the parents broke down, but the children remained with Ms R.I.
- R. had unrestricted access to the children and played an active role in their education.

..... **(total 494 tokens)**.....

- R. informed the authorities that he had moved with the children to Năvodari, Constanța County.
- The Constanta District Court partially allowed R.'s opposition to the enforcement, finding that he had complied in full with the obligation to provide for the children since they lived with him.
- The children refused to move in with their mother in a meeting at the bailiff's office.
- Ms R.I. lodged a criminal complaint against R., accusing him of disobeying the court order concerning the custody arrangements.

Figure 5: An example input and output of the LLM about data 001-187931. The original document has 3618 tokens totally. It reduces to 494 tokens after extracting important points of a legal case.