Rethinking the Development of Large Language Models from the Causal Perspective: A Legal Text Prediction Case Study

Haotian Chen[1], **Lingwei Zhang**[2], **Yiran Liu**[3], **Yang Yu**[3]

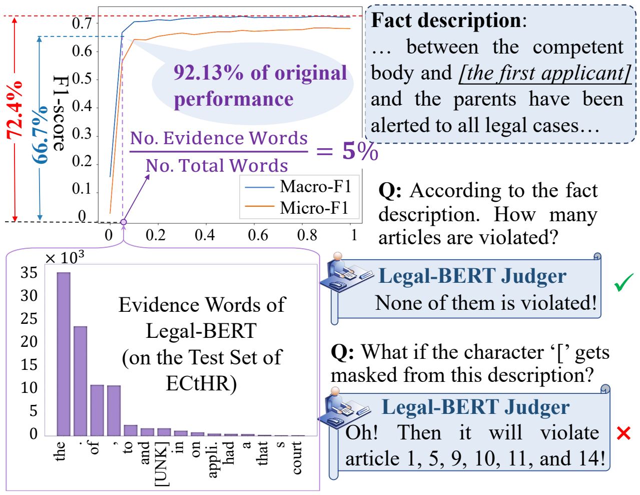
[1]School of Computer Science, Fudan University, Shanghai, China
[2]Department of Computer Science, Johns Hopkins University, Baltimore, USA
[3]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
htchen18@fudan.edu.com, lzhan218@jhu.edu, liu-yr21@mails.tsinghua.edu.cn, yangyu1@tsinghua.edu.cn

Abstract

While large language models (LLMs) exhibit impressive performance on a wide range of NLP tasks, most of them fail to learn the causality from correlation, which disables them from learning rationales for predicting. Rethinking the whole developing process of LLMs is of great urgency as they are adopted in various critical tasks that need rationales, including legal text prediction (e.g., legal judgment prediction). In this paper, we first explain the underlying theoretical mechanism of their failure and argue that both the data imbalance and the omission of causality in model design and selection render the current training-testing paradigm failed to select the unique causality-based model from correlation-based models. Second, we take the legal text prediction task as the testbed and reconstruct the developing process of LLMs by simultaneously infusing causality into model architectures and organizing causality-based adversarial attacks for evaluation. Specifically, we base our reconstruction on our theoretical analysis and propose a causality-aware self-attention mechanism (CASAM), which prevents LLMs from entangling causal and non-causal information by restricting the interaction between causal and non-causal words. Meanwhile, we propose eight kinds of legal-specific attacks to form causality-based model selection. Our extensive experimental results demonstrate that our proposed CASAM achieves state-of-the-art (SOTA) performances and the strongest robustness on three commonly used legal text prediction benchmarks. We make our code publicly available at https://github.com/Carrot-Red/Rethink-LLM-development.

1 Introduction

Large language models (LLMs) have undergone significant development and significantly impacted our life in a wide range of applications (OpenAI 2023), including legal judgment prediction (Feng, Li, and Ng 2022; Chalkidis et al. 2022a), drug discovery (Singha Roy and Mercer 2023), and quantitative trading (Sawhney et al. 2021; Ju et al. 2023). While we enjoy their human-surpassing performance, they exhibit certain risks caused by confusing causality from correlation (Chen, Chen, and Zhou 2023). Their failure of learning causality (rationales) not only degrades their performance but also renders them untrustworthy, thus impeding their real-world applications, especially in those high-stake tasks that require rationales for decisions (e.g., legal judgment prediction). For example, we found clues that the commonly adopted state-of-the-art (SOTA) LJP models (Chalkidis et al. 2020; Zheng et al. 2021; Chalkidis et al. 2022a) learn the spurious correlations and simultaneously achieve the minimum training loss. As shown in Figure 1, the prediction of Legal-BERT (Chalkidis et al. 2020) is reversed by a small change that does not cause an essential semantic change. Furthermore, the most important keywords deciding the model predictions mainly concentrate on punctuation marks and function words. A large number of predictions only rely on less than 5% of words from the fact descriptions rather than considering the whole text. These potential risks make the causality-understanding model become an urgent need. We have to rethink the whole developing process of LLMs.



Fact description: … between the competent body and *[the first applicant]* and the parents have been alerted to all legal cases…

Q: According to the fact description. How many articles are violated?

None of them is violated!

Q: What if the character '[' gets masked from this description?

Oh! Then it will violate article 1, 5, 9, 10, 11, and 14!

Figure 1: An example of reversed prediction caused by character substitution. We also present the frequency of evidence words considered by Legal-BERT. Making predictions merely according to these evidence words (5% words of each input document) reserves 92.13% performance of Legal-BERT. "No." denotes the number of words.

In this paper, we first argue that data imbalance is the main factor that hampers the traditional developing paradigm from obtaining a causality-based model. Both the uniq causality-based model and spurious-correlation-based mo els are able to achieve competitive or state-of-the-art (SOT performance in traditional datasets due to the inevitable d imbalance, which is caused by the sampling process a linguistic rules. For example, function words like ‘a’ occ more frequently than many other words. Since data-driv AI models are tied to datasets where training and test d are assumed to be identically and independently distribut (*i.i.d.* assumption), spurious correlations in the training d are also established in the test data. Therefore, spuriou correlation-based models are able to achieve SOTA perf mance. The current training-testing paradigm, focusing chasing the SOTA performance, fails to train and select t unique causality-based model and distinguish causality from correlation.

Second, we take a step further toward revealing two long-standing omissions of previous SOTA-chasing methods from the causal perspective. According to our theoretical analysis, the two omissions exacerbate the confusion of learning models between causality and correlation under the data imbalance circumstance. The first omission is that the learning models miss the guidance of causality (e.g., the use of human knowledge), which largely increases their uncertainty when inferring causal relationships from the training data. For example, if we tell a linear model to zero the weight of all non-causal variables, the model will possess a strong ability to learn causality despite the data imbalance. The second omission is that most current evaluation methods focus on measuring average error across a held-out test set instead of evaluating the causality-understanding ability of models. Models just need to greedily absorb all correlations that happen to be predictive in the test set even if they are not causal relationships.

Third, to address the issue, we reconstruct the whole developing process of LLM and take the legal text prediction tasks as our testbed to verify the effectiveness of our reconstruction. Our reconstruction includes infusing causality into the architecture of models and evaluating their causality-understanding ability. Specifically, to infuse causal knowledge into learning models, we aim to prevent them from learning non-causal information by restricting the interaction (represented by attention weights) between causal and non-causal words. We propose a causality-aware self-attention mechanism (CASAM) to reallocate the attention weights throughout the overall transformer encoder, which leads the LLM to pay more attention to causal information. Meanwhile, we accurately select the causality-based model by adopting our proposed testing method named causality-based adversarial attacks consisting of eight kinds of attacks. Models can pass the evaluation only when they successfully learn stable causal relationships. Otherwise, their performances drop sharply as spurious correlations are not established in adversarial samples. The extensive experimental results show that our proposed CASAM performs better on generalization and robustness than the baseline models and achieves new SOTA performance on the three commonly used legal prediction datasets.

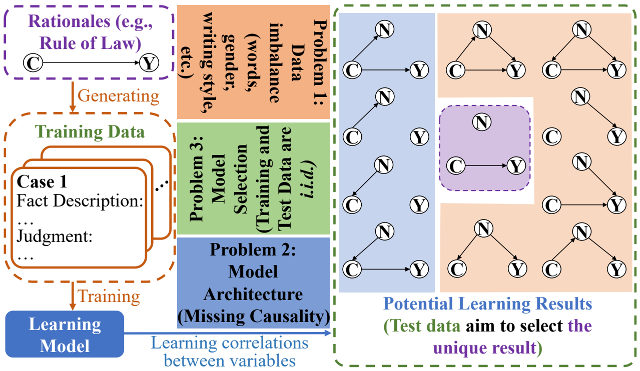

Rationales (e.g., Rule of Law)
C
Y
Generating
Training Data
Case 1
Fact Description:
...
Judgment:
...
Training
Learning Model
Learning correlations between variables
Problem 1: Data imbalance (words, gender, writing style, etc.)
Problem 3: Model Selection (Training and Test Data are i.i.d.)
Problem 2: Model Architecture (Missing Causality)
N
C
Y
Potential Learning Results
(Test data aim to select the unique result)

Figure 2: Problem illustration. C, N, and Y denote causal information, non-causal information, and final prediction, respectively. Rationales (causality) generate training data, and models are expected to accurately estimate the correlations between variables (causal relationships in purple). Multiple potential results are all optimal and possible to be learned by models. Data augmentation (in orange), merely controlling the effect of N on Y, filters the six orange graphs. A causality-aware learning model (in blue) filters ten graphs, including all the orange and blue ones.

2 Failure of Learning Causality

In this section, we point out three main problems, namely data imbalance, missing causality in model architecture, and missing causality in model selection, that hamper the previous LLM-developing methods from both training and selecting a causality-based model. The problems and the nature of the methods are depicted in Figure 2: Data imbalance makes multiple learning results of the same model architecture possible to be optimal and succeed in passing the evaluation stage. We elaborate on this issue and the underlying causes as follows.

Data Imbalance. Data imbalance actually distorts the information adopted to train and test models. We use data-driven methods to train different models for learning the correlation relationships between input variables and output variables (Cui and Athey 2022). The learned correlations among variables can be generated in either of the three ways: causality, confounding, and data selection bias (Cui and Athey 2022). Only the correlations generated by causality are what we expect the models to learn from. However, both natural language and social unfairness make the raw data of case descriptions compose an imbalanced dataset of C, N, and Y, which exacerbates the data selection bias (data imbalance) and thus leads to the correlations generated by it. Such correlations are referred to as spurious correlations. Learning models greedily absorb all correlations to minimize the training loss (Ye et al. 2021), which renders them spurious-correlation-based models.

Multiple Learning Results. It becomes a random event R whether our selected model trained by the imbalanced data is the unique causality-based model (which learns the correlation generated by $C \rightarrow Y$). The randomness of R stems from the stochastic learning process and the number of spurious correlations established in the test data, which leads to a certain probability distribution across all potential learning (estimating) results shown on the right of Figure 2. They are possible to be optimal as they can experimentally and theoretically minimize the loss function to zero by absorbing all correlations. *Note that different Learning results present distinctive decision rules of models (e.g., making legal judgments according to gender).*

Missing Causality in Model Architecture. Without prior knowledge containing causal information, models merely learn the correlations between variables without distinguishing causal information from non-causal information. In that case, the spurious correlations generated by data imbalance are inevitably learned by models. As shown in Figure 2, the four blue causal graphs remain possible for becoming the decision rules of learning models.

Missing Causality Evaluation in Model Selection. By learning spurious correlations, models can succeed in making the right predictions as causal-based models do due to the *i.i.d.* assumption: the learned spurious correlations in training data will also be predictive in testing data. Therefore, it is of great essence to design causality-specific evaluation methods for selecting the unique causality-based model.

3 Redesigning the Developing Process of LM

We posit three solutions to address the issue mentioned in Section 2 and adopt the latter two to propose our method.

Data Governance. We can constrict the effect of N on Y by data governance (e.g., data augmentation). However, as shown in Figure 2, such kind of method only filters 6 out of 10 potential learning results without interrupting the correlation between C and N, which suppresses the probability of R.

Infusing Causality into Model. We improve the understanding ability of models, making them able to distinguish and avoid learning spurious correlations. For example, a straightforward linear model can avoid the disturbance of non-causal information if it knows the legal knowledge: It can set the coefficient in front of non-causal variables to zero regardless of their amount in the training data. To this end, we propose CASAM for infusing learning models with causal information and knowledge.

Causality-Invariant Attacks for Evaluation. We complete the testing data by proposing legal-specific attacks to bring distributional shifts into the data. According to the suggestions provided by experts in the legal domain, we consider several types of attacks for thorough robustness evaluation. In each type of the following attacks written in **bold**, we make a distinct perturbation in the given fact description that will not change the judgment from the perspective of the experts. For those attacks written in *italics*, the perturbation will not change the judgment in most circumstances according to the experts. We provide descriptions of all types of attacks: (1) *functional word attacks*. We adopt the token '[mask]' as a substitute for a functional word; (2) *word-level attacks*, which mask a single word; (3) **sequence number attacks**, which remove the sequence number in front of the given description; (4) **dot attacks after sequence number**. We remove the dot after a sequence number; (5) **punctuation mark attacks**, which mask a punctuation mark; (6) **auxiliary verb attacks**, which mask an auxiliary verb; (7) **article attacks**, which mask an article before a noun; (8) **preposition attacks**. We attack prepositions except for the preposition 'of' (which may indicate the ownership relationship), the preposition 'for' (which may represent whether someone does something on purpose), and those prepositions that locate between numbers.

4 Methodology

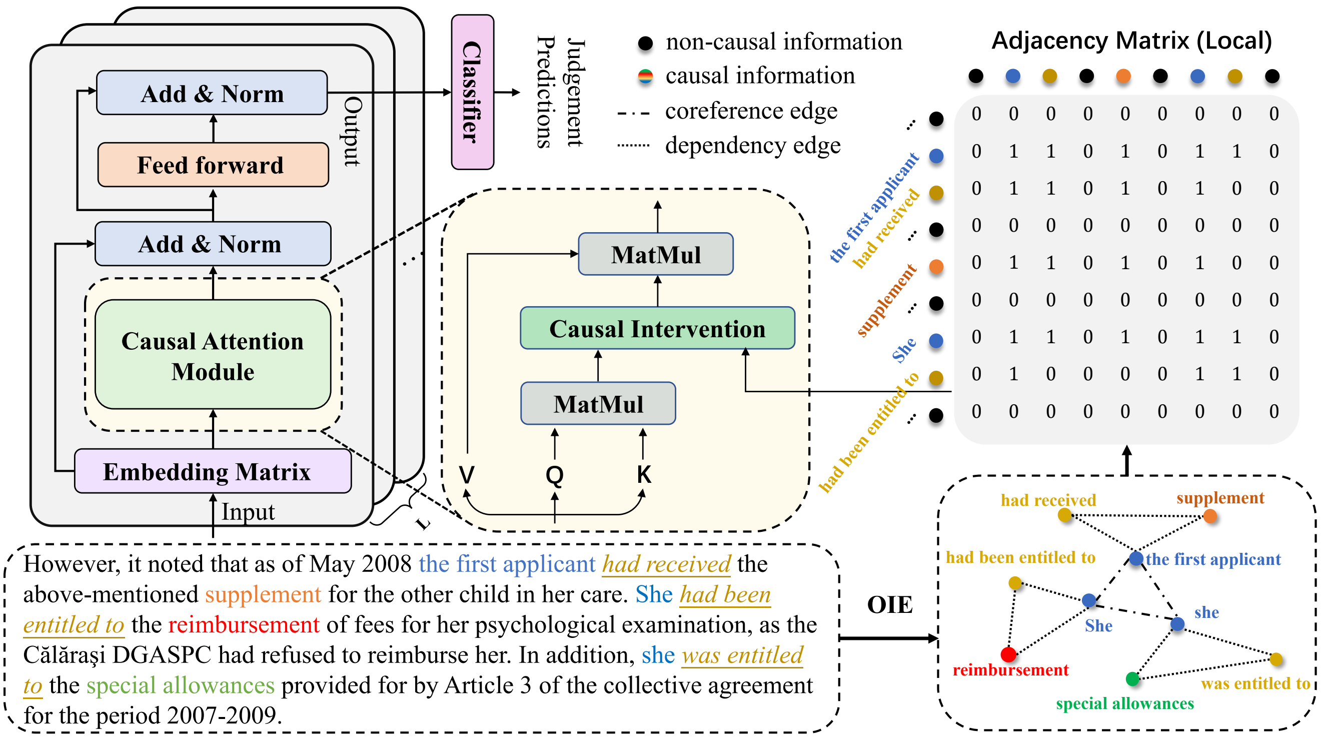
Our overall framework is shown in Figure 3 and can be divided into two steps. In the first step, we adopt the OIE and open-source coreference methods to refine the dataset and mitigate the data imbalance in legal texts. We first perform open information extraction (OIE) on input legal texts to discard the context that contains a high proportion of non-causal information. Then, we graphically structure the extracted pieces of information. In the extracted information (knowledge) graph, the nodes denote the subjects, objects, and predicates while the edges are dependencies. The nodes possessing the same semantic meaning will be merged into one by the open-source coreference model. During the process of constructing graphs, redundant non-causal information is further reduced by merging. Meanwhile, documents are substantially compressed to focus on core information. The above data processing reduces the information entropy of distinguishing N and C. In the second step, we apply the knowledge to intervene in the learning process. In the rest of this section, we provide the detail of our methods.

4.1 Graph Construction by OIE

We adopt OIE aiming to discard and merge redundant non-causal information. First, we apply coreference resolution (Clark and Manning 2016) and open information extraction (Stanovsky et al. 2018) tools to identify the corresponding mentions or pronouns of each entity, and then extract relational triplets from sentences. In our constructed graph, we represent subjects and objects as nodes, which are connected by predicates as directed edges. Second, the nodes will be merged to reduce redundant non-causal information if they have similar names or meanings, which is identified by TF-IDF overlap and coreference resolution tools, respectively. Finally, as to the subsequent newly extracted triplets, we also calculate the TF-IDF overlap between the existing triplets and the new one. If the value is higher than our predefined threshold, we rule out the new triplet to reduce information replication.

4.2 CASAM

We introduce our proposed CASAM in this section. CASAM partly inherits the architecture of the transformer (Vaswani et al. 2017) or Legal-BERT (Chalkidis et al. 2020) encoder which consists of L stacking blocks. Each block comprises a feed-forward network, residual connection, layer normalization, and a causal attention module. Given a fact description D, we obtain its embedding matrix



	●	●	●	●	●	●	●	●	●
●	0	0	0	0	0	0	0	0	0
the first applicant ●	0	1	1	0	1	0	1	1	0
had received ●	0	1	1	0	1	0	1	0	0
●	0	0	0	0	0	0	0	0	0
supplement ●	0	1	1	0	1	0	1	0	0
●	0	0	0	0	0	0	0	0	0
She ●	0	1	1	0	1	0	1	1	0
had been entitled to ●	0	1	0	0	0	0	1	1	0
●	0	0	0	0	0	0	0	0	0

However, it noted that as of May 2008 the first applicant had received the above-mentioned supplement for the other child in her care. She had been entitled to the reimbursement of fees for her psychological examination, as the Călăraşi DGASPC had refused to reimburse her. In addition, she was entitled to the special allowances provided for by Article 3 of the collective agreement for the period 2007-2009.

Figure 3: Overview of our framework.

$\mathbf{X} \in \mathbb{R}^{N\times d}$ according to the embedding layer of a transformer encoder, where N and d denote the sequence length and the dimension of hidden layers, respectively. Following the transformer encoder, CASAM maps $\mathbf{X}$ to query, key, and value matrices in each block by $\mathbf{Q} = \mathbf{XW}_q, \mathbf{K} = \mathbf{XW}_k, \mathbf{V} = \mathbf{XW}_v$, where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d\times k}$ are model parameters the in l-th block, l is omitted in the equation for brevity.

Different from the widely adopted self-attention mechanism which considers all words to correlate with each other, our proposed causal attention module performs causal intervention between each word pair. The former provides abundant correlations represented by unsupervised attention weights for models to explore, neglecting the fact that learning methods will greedily absorb all correlations (including spurious correlations) found in data to minimize their training error, which leads to spurious correlation error. The latter tries to discern the potential causal relationships and block non-causal information to prevent learning spurious correlations. Specifically, our proposed CASAM first derives an adjacency matrix $\mathbf{A}$ according to a certain graph $\mathcal{G}$ constructed by the aforementioned open information extraction (OIE) tool. The entries $\mathbf{A}_{ij}$ tabulate the binary variable identifying whether the combination of i-th word and j-th word will causally affect the final judgment. Then, based on the original attention weights calculated by $\mathbf{S} = \frac{\mathbf{QK}^T}{\sqrt{d}}$, the new attention weights are derived by,

$$\mathbf{S}' = \alpha\mathbf{S} + (1-\alpha)\mathbf{S} \odot \mathbf{A}, \tag{1}$$

where $\odot$ denotes the element-wise multiplication between matrices and α is a hyperparameter ranging from 0 to 1, which is adjusted according to the accuracy of an OIE tool: the more accurate the OIE tool, the higher the α. The output $\mathbf{Y}$ of each causal attention module is derived by,

$$\mathbf{Y} = \mathrm{softmax}(\mathbf{S}')\mathbf{V}. \tag{2}$$

We input $\mathbf{Y}$, the output of the final causal attention layer considered as the representation of a fact description D, into a linear layer followed by a sigmoid function to obtain the final predictions.

4.3 Model Selection

Traditionally, training data and validation data are *i.i.d.*, and validation data are adopted to monitor the training process for selecting the best-performed version of a learning model. According to our analysis in Section 3, the selection can be biased as the evaluation of the generalization ability of models is incomplete: lacking the evaluation of out-of-distribution (OOD) performance. To solve the issue, we complete the validation data by our proposed legal-specific attacks to evaluate both the robustness and generalization ability of models. Different from previous methods, we aim to select the most robust and generalizable version of a learning model during the training process.

5 Experiments

5.1 Datasets

ECtHR Task A & B. The European Court of Human Rights (ECtHR) dataset (Chalkidis, Androutsopoulos, and Aletras

2019) is the only publicly available human-annotated LJP dataset in English, consisting of approximately 11,000 cases from the ECtHR database. In each case, allegations are written as fact descriptions, the judgment results — about which of human rights provisions legislated by European Convention of Human Rights (ECHR) does the current state breach — are recorded as the label. All cases are chronologically categorized as training set (9k, 2001-2016), development set (1k, 2016-2017), and test set (1k, 2017-2019). Each case can either violate single, multiple, or none of the given legal articles. For each model, the input is fact descriptions of a case, and the output is the judgment, represented by a set of violated articles. In Task A, the violated articles are considered by the court. In Task B, the violated articles are put forward by the applicants.

LEDGAR. LEDGAR (Labeled EDGAR) (Tuggener et al. 2020) is a dataset for contract provision classification. Considering the underlying common legal text classification techniques, we conduct experiments on the dataset not only for a more comprehensive evaluation, but also to test the generalization ability of models. In LEDGAR, the contract provisions are crawled from the U.S. Securities and Exchange Commission (SEC) website and are available from an Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system on the website. Nearly 850k contract provisions from 12.5k categories are included in the originally proposed LEDGAR. Following the legal language understanding benchmark LexGLUE (Chalkidis et al. 2022a), we use 80k contract provisions labeled with 100 most frequent categories from the original dataset. The dataset is chronologically split into a training set (60k, 2016-2017), a development set (10k, 2018), and a test set (10k, 2019).

5.2 Baselines

Following the previous methods in legal text prediction (Chalkidis et al. 2022a), we compare our proposed CASAM with the eight baseline models: TFIDF+SVM, BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), DeBERTa (He et al. 2021), Longformer (Beltagy, Peters, and Cohan 2020), BigBird (Zaheer et al. 2020), CaseLaw-BERT (Zheng et al. 2021), and Legal-BERT (Chalkidis et al. 2020).

The backbone of our proposed model is based on Legal-BERT in this paper, our model can be easily extended to other backbones in future work.

5.3 Experimental Settings

Implementation Details. Our experiment is based on PyTorch and Hugging Face Transformer (Wolf et al. 2020). At the graph construction stage, co-reference resolution predictor (Clark and Manning 2016) and OIE predictor (Stanovsky et al. 2018) are used to extract graph relationships and construct the graph. Later, we use breadth-first search to get the linearized graph text. We apply the pre-trained Legal BERT transformer from Hugging face to be our encoder. With the original fact descriptions and the corresponding graph text, we use two Legal BERT encoders to get the embeddings. The learning rate is $1e-4$ and the optimizer is AdamW.

Method	ECtHR(A)		ECtHR(B)		LEDGAR	
	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1
TFIDF+SVM*	64.5	51.7	74.6	65.1	87.2	82.4
BERT	71.1	61.2	79.2	72.1	88.0	82.1
RoBERTa	72.0	65.6	77.6	70.9	87.6	81.3
DeBERTa	71.5	66.7	80.2	73.1	88.1	82.9
Longformer	71.0	62.1	79.7	71.9	87.7	81.3
BigBird	69.8	59.7	78.1	68.5	87.1	80.8
Legal-BERT	72.3	66.0	80.6	75.2	88.2	81.9
CaseLaw-BERT	71.6	65.5	78.6	71.9	88.0	81.8
CASAM	**73.8**	**68.5**	**81.4**	**76.0**	**88.7**	**83.5**
w/o causal attention	72.3	66.0	80.6	75.2	88.2	81.9

Table 1: Overall experimental results. The signal '*' denotes that the results of the corresponding models are quoted from LexGLUE (Chalkidis et al. 2022a).

Evaluation Metrics. To evaluate the robustness of models, we adopt **certified ratio** (Gürel et al. 2022), namely CR, to measure the percentage of consistent predictions (unchanged predictions) under a perturbation (wrong predictions are also included). Following previous work (Chalkidis et al. 2022a), we evaluate the performance (e.g., generalization ability) of models by μ-F_1 and m-F_1 scores.

Attribution Method. Current feature attribution methods can be roughly divided into three categories: gradient-based methods which calculate a score for each input feature by gradients (Springenberg et al. 2015; Li et al. 2016; Simonyan, Vedaldi, and Zisserman 2014), reference-based methods which consider the difference between a predefined "reference" and the output of a model as the attribution score (Ribeiro, Singh, and Guestrin 2016; Shrikumar, Greenside, and Kundaje 2017; Sundararajan, Taly, and Yan 2017), and erasure-based methods which measure the change of model prediction as the attribution score after removing the target feature (Zeiler and Fergus 2014; Li et al. 2016; Feng et al. 2018; Chen, Zheng, and Ji 2020). We adopt an erasure-based method (Li et al. 2020) due to its simplicity and faithfulness. Specifically, if a fact description $D = [d_1, \ldots, d_{i-1}, d_i, d_{i+1}, \ldots]$ is input into a certain model, and the corresponding output prediction score on the ground truth label y is $o_y(D)$, then the attribution value on d_i is written by,

$$F_y(d_i) = o_y(D) - o_y(D'), \tag{3}$$

where $D' = [d_1, \ldots, d_{i-1}, [\text{MASK}], d_{i+1}, \ldots]$. Erasure-based methods directly satisfy the way of evaluating an AI judger by the rule of law: Will the judgment change if the causal elements get erased or changed from the fact descriptions? Will the AI judger consistently stick to the rule of law in any circumstances (e.g., changes in irrelevant information)?

5.4 Main Results and Ablation Study

The generalization ability (performance) evaluation results of baselines and our model are shown in Table 1. We can observe that the performance of our CASAM significantly outperforms the SOTA baseline methods, achieving

Method	Attack 1			Attack 2			Attack 3		Attack 4		Attack 5			Attack 6			Attack 7			Attack 8		
	E.A	E.B	L.	E.A	E.B	L.	E.A	E.B	E.A	E.B	E.A	E.B	L.	E.A	E.B	L.	E.A	E.B	L.	E.A	E.B	L.
BERT	99.4	99.3	93.8	99.3	99.2	84.0	99.4	99.4	99.6	99.4	99.4	99.4	96.5	99.5	99.2	98.2	99.4	99.3	98.5	99.4	99.3	93.8
RoBERTa	99.3	99.1	95.8	99.2	98.9	79.1	99.7	99.5	99.4	99.3	99.3	99.1	96.0	99.0	99.0	70.0	99.0	98.9	98.4	99.3	99.1	95.6
Longformer	91.5	93.6	**96.5**	81.9	86.1	78.9	-	-	-	-	91.6	93.6	96.8	99.1	98.2	90.0	96.5	97.5	98.1	91.5	93.6	**96.4**
Bigbird	96.1	95.8	96.4	96.1	95.8	96.4	-	-	-	-	96.4	96.0	96.7	99.3	99.3	90.0	98.9	99.1	97.4	96.2	95.8	96.4
CaseLaw	99.6	99.6	94.3	99.5	99.5	94.3	99.5	99.5	99.7	99.7	99.6	99.6	97.5	99.6	99.6	**99.0**	99.6	99.6	98.5	99.5	99.5	94.4
Legal-BERT	99.8	99.7	94.8	99.6	99.5	88.7	**99.8**	99.6	**99.8**	**99.9**	99.8	99.6	96.8	99.8	99.6	96.8	99.7	99.7	98.4	99.7	99.7	94.5
CASAM	**99.9**	**99.9**	96.2	**99.8**	**99.8**	**89.7**	**99.8**	**99.9**	**99.8**	**99.9**	**99.8**	**99.9**	**98.1**	**99.9**	**99.9**	98.9	**99.9**	**99.9**	**99.2**	**99.9**	**99.9**	96.3

Table 2: Results of robustness evaluation measured by certified ratio (CR) on the test sets of three benchmark datasets. "E.A", "E.B", and "L." denote dataset ECtHR Task A, ECtHR Task B, and LEDGAR, respectively. Details of each kind of attack are introduced in Section 3.

a new SOTA performance on all three benchmark datasets. Note that our framework is based on the Legal-BERT backbone. Compared with Legal-BERT, CASAM yields gains of 3.8%/4.5% of μ/m-F1 scores in ECtHR Task A, 1.0%/1.3% of μ/m-F1 scores in ECtHR Task B, and 0.5%/0.5% of μ/m-F1 scores in LEDGAR. The experimental results in Table 1 indicate that, with the guidance of our theoretical analysis, CASAM effectively improves the performance of Legal-BERT: it blocks N to reduce the spurious correlation error, which leads the model to learn the underlying ground-truth knowledge and thereby enhancing the generalization ability of the model.

5.5 Results of Robustness Evaluation

We evaluate the robustness of models against diverse attacks. As shown in Table 2, the robustness of our proposed CASAM is significantly stronger than its backbone on the three datasets under all kinds of attacks. Without any modification, the original Legal-BERT exhibits poor robustness, especially on LEGDAR. Changes in the irrelevant information in fact descriptions will eventually render the Legal-BERT judger altering at most 11.27% of its predictions, which terribly hurts its robustness and trust in it. Such kinds of mistakes caused by the spurious correlation error impede the deployment of AI judgers in real-world applications. Our proposed CASAM significantly mitigates the underlying error and thus enhances the robustness of models. In the three legal text prediction tasks, our proposed CASAM achieves the certified ratio over 99%, which indicates that it gets extremely close to the standard of being trustworthy under diverse attacks proposed by experts in the legal domain.

We can observe that CASAM achieves close performance on judgment prediction tasks. We posit the underlying reason: CASAM satisfies the common theoretical background, intervening in the architecture of the model and breaking the correlation between N and C in the training procedure, thereby preventing N from correlating with Y. As we mentioned in Section 2 that only the correlations generated by causality are what we expect the models to learn from, CASAM focuses on removing other kinds of correlations and only reserving those generated by causality.

Despite the significant robustness improvement under all kinds of attacks, we explain the reason why the evaluation results of our proposed methods under word-level attacks are largely different from other attacks on LEDGAR. Although legal judgment prediction and legal text classification often share common techniques, the underlying decision rules of the two task is different. Different from LJP where a judger is required to both perform legal reasoning and consider all of the circumstances in a case for a just judgment, we rely on fewer words in legal text classification. For example, if we notice the word 'vegetables', 'fruits', or 'agriculture' in a legal file, we know it probably belongs to the 'agriculture' category. If we mask these words, it will even be difficult for humans to classify the file. Under word-level attacks, these words will inevitably be masked, leading to distinct evaluation results.

5.6 Analysis and Discussion

In this section, we take a step further toward characterizing both the data imbalance and the decision rules (evidence words) of models in the context of the legal text prediction task. We shed some light on the underlying reasons why our proposed methods achieve stronger generalization ability and robustness.

Visualization of Selection Bias. To characterize the data imbalance (selection bias), we investigate the ECtHR Task A dataset as an example and analysis to what extent the Legal-BERT is affected by the bias. First, we use a feature attribution method to obtain the top 5% words considered most crucial by Legal-BERT when making a judgment prediction in the test set. Second, we count the frequency of each word in the top 5% words and in the training set of ECtHR Task A, respectively. As shown on the left of Figure 4, we can observe three phenomena: (1) there is an obvious word frequency bias in the training set of ECtHR Task A; (2) the same kind of bias occurs in the top 5% crucial words considered by Legal-BERT; (3) the two kinds of frequency exhibit a common distribution. The first phenomenon, exhibiting a severe bias in the training set, can lead learning models to suffer from data imbalance, which is demonstrated by the causal structural model in Figure 2 and instantiated by the second phenomenon. The third phenomenon indicates the fact that, without any intervention, learning models will faithfully learn the bias distribution in the training data, which is undesirable for all heuristic learning methods. If

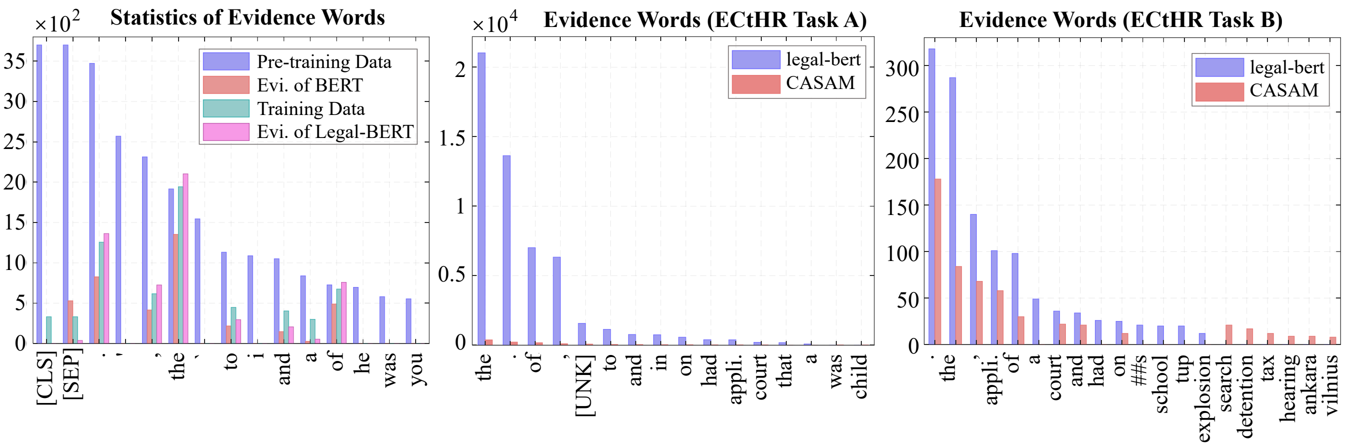


Figure 4: Visualization of the frequency of evidence words. "Evi.", "Pre-training data", and "Training data" denote the evidence words, the legal corpus used to pre-train BERT, and the training data of ECtHR Task A for finetuning, respectively.

the bias brought by a training set is correlated with gender, race, or geography, learning models will even trigger severe social problems.

Effect of Debiasing. To investigate the reason why our proposed methods possess better generalization ability and robustness, we visualize their decision rules of predicting judgment in the test set of ECtHR A and B. We count the frequency of each word occurring in the top 5% words, which are considered most crucial by Legal-BERT and our proposed CASAM, respectively. The results are shown on the middle and right of Figure 4. Our observations are threefold: (1) Without any adjustments in training data or the architecture of the model, Legal-BERT significantly correlates non-causal information with the judgments. It predicts judgments through those words that hardly possess any semantic meanings. The spurious correlations render Legal-BERT vulnerable to attacks and impede its deployment in real-world legal scenarios: The changes in non-causal information like writing style (frequently or rarely use these function words) can even affect the predictions of Legal-BERT. (2) After our intervention in the architecture of Legal-BERT, it significantly decouples the non-causal information (e.g., the punctuation marks and function words) and final predictions, which presents the effectiveness of reducing the possibility of potential learning results shown in Figure 2. (3) Our intervention makes Legal-BERT learn new causal information (e.g., content words that indeed affect the predictions), especially in the middle of Figure 4, which indicates that our proposed methods succeed in learning causal information (the ground-truth estimate) for predicting by reducing the possibility of other potential estimates. This explains why our proposed methods achieve both SOTA generalization ability and robustness.

Note that CASAM can still be aware of non-causal information in some situations shown in the right of Figure 4 due to the precision of OIE tools: N cannot be precisely distinguished from C, which hampers more performance gains of our proposed methods. We leave the improvement of OIE tools for future work.

6 Related Work

The rapid development of large-scale pre-trained language models (PLMs) based on transformers significantly benefits a wide range of downstream tasks such as legal text processing (Cui et al. 2022). Some of the PLMs including BERT (Devlin et al. 2018) are further pre-trained on domain-specific corpora, such as Legal-BERT (Chalkidis et al. 2020) which exhibits the SOTA performance on legal text processing benchmarks (e.g., LexGLUE) (Zheng et al. 2021; Chalkidis et al. 2022a). However, in the meantime, some severe problems of models are also discovered, including unfairness and discrimination (Chalkidis et al. 2022b). Accordingly, researchers propose debiasing methods (Guo, Yang, and Abbasi 2022; Sevim, Şahinuç, and Koç 2022) to mitigate the bias or conduct substantial experiments to investigate and analyze the decision rules of PLMs (Clark et al. 2019; Chen, Chen, and Zhou 2023). Different from previous work, we rethink the development of LLMs from the causal perspective to theoretically analyze the underlying causes of their problems. After that, we give our solution and finally demonstrate its effectiveness through extensive experimental results.

7 Conclusion

In this paper, we investigate the decision rule of the legal-specific PLM in legal AI. We exhibit the potential problems of the decision rules caused by spurious correlation error and propose a structural causal model to theoretically analyze the underlying mechanism. Under the guidance of our analysis, we propose a method to simultaneously reduce non-causal information and retain causal information in the given fact descriptions. The experimental results indicate that spurious correlations between non-causal information and predictions largely damage the generalization ability and robustness of legal AI. We appeal to future work to take the spurious correlation error into consideration for improving the overall performance of legal AI.

References

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. *ArXiv*.

Chalkidis, I.; Androutsopoulos, I.; and Aletras, N. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4317–4323. Florence, Italy: Association for Computational Linguistics.

Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; and Androutsopoulos, I. 2020. LEGAL-BERT: The Muppets Straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2898–2904. Online: Association for Computational Linguistics.

Chalkidis, I.; Jana, A.; Hartung, D.; Bommarito, M.; Androutsopoulos, I.; Katz, D.; and Aletras, N. 2022a. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4310–4330. Dublin, Ireland: Association for Computational Linguistics.

Chalkidis, I.; Pasini, T.; Zhang, S.; Tomada, L.; Schwemer, S.; and Søgaard, A. 2022b. FairLex: A Multilingual Benchmark for Evaluating Fairness in Legal Text Processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4389–4406. Dublin, Ireland: Association for Computational Linguistics.

Chen, H.; Chen, B.; and Zhou, X. 2023. Did the Models Understand Documents? Benchmarking Models for Language Understanding in Document-Level Relation Extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6418–6435. Toronto, Canada: Association for Computational Linguistics.

Chen, H.; Zheng, G.; and Ji, Y. 2020. Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5578–5593. Online: Association for Computational Linguistics.

Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276–286. Florence, Italy: Association for Computational Linguistics.

Clark, K.; and Manning, C. D. 2016. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2256–2262. Austin, Texas: Association for Computational Linguistics.

Cui, J.; Shen, X.; Nie, F.; Wang, Z.; Wang, J.; and Chen, Y. 2022. A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges. arxiv:2204.04859.

Cui, P.; and Athey, S. 2022. Stable Learning Establishes Some Common Ground between Causal Inference and Machine Learning. *Nature Machine Intelligence*, 4(2): 110–115.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. N. 2018. 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)*, 4171–4186.

Feng, S.; Wallace, E.; Grissom II, A.; Iyyer, M.; Rodriguez, P.; and Boyd-Graber, J. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3719–3728. Brussels, Belgium: Association for Computational Linguistics.

Feng, Y.; Li, C.; and Ng, V. 2022. Legal Judgment Prediction: A Survey of the State of the Art. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 5461–5469. Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-00-3.

Guo, Y.; Yang, Y.; and Abbasi, A. 2022. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1012–1023. Dublin, Ireland: Association for Computational Linguistics.

Gürel, N. M.; Qi, X.; Rimanic, L.; Zhang, C.; and Li, B. 2022. Knowledge Enhanced Machine Learning Pipeline against Diverse Adversarial Attacks. arxiv:2106.06235.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Ju, J.-H.; Huang, Y.-S.; Lin, C.-W.; Lin, C.; and Wang, C.-J. 2023. A Compare-and-contrast Multistage Pipeline for Uncovering Financial Signals in Financial Reports. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14307–14321. Toronto, Canada: Association for Computational Linguistics.

Li, J.; Chen, X.; Hovy, E.; and Jurafsky, D. 2016. Visualizing and Understanding Neural Models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 681–691. San Diego, California: Association for Computational Linguistics.

Li, L.; Ma, R.; Guo, Q.; Xue, X.; and Qiu, X. 2020. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6193–6202. Online: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A Robustly Optimized Bert Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

OpenAI. 2023. GPT-4 Technical Report. arxiv:2303.08774.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. San Francisco California USA: ACM. ISBN 978-1-4503-4232-2.

Sawhney, R.; Wadhwa, A.; Agarwal, S.; and Shah, R. R. 2021. Quantitative Day Trading from Natural Language Using Reinforcement Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4018–4030. Online: Association for Computational Linguistics.

Sevim, N.; Şahinuç, F.; and Koç, A. 2022. Gender Bias in Legal Corpora and Debiasing It. *Natural Language Engineering*, 1–34.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning Important Features through Propagating Activation Differences. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 3145–3153. PMLR.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Proceedings of the International Conference on Learning Representations (ICLR), 1–8. ICLR.

Singha Roy, S.; and Mercer, R. E. 2023. Extracting Drug-Drug and Protein-Protein Interactions from Text Using a Continuous Update of Tree-Transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, 280–291. Toronto, Canada: Association for Computational Linguistics.

Springenberg, J.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2015. Striving for Simplicity: The All Convolutional Net. In *ICLR (Workshop Track)*.

Stanovsky, G.; Michael, J.; Zettlemoyer, L.; and Dagan, I. 2018. Supervised Open Information Extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 885–895. New Orleans, Louisiana: Association for Computational Linguistics.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 3319–3328. PMLR.

Tuggener, D.; von Däniken, P.; Peetz, T.; and Cieliebak, M. 2020. LEDGAR: A Large-Scale Multi-Label Corpus for Text Classification of Legal Provisions in Contracts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 1235–1241. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, 5998–6008.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Ye, H.; Xie, C.; Cai, T.; Li, R.; Li, Z.; and Wang, L. 2021. Towards a Theoretical Framework of Out-of-Distribution Generalization. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 23519–23531. Curran Associates, Inc.

Zaheer, M.; Guruganesh, G.; Dubey, K. A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; and Ahmed, A. 2020. Big Bird: Transformers for Longer Sequences. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 17283–17297. Curran Associates, Inc.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and Understanding Convolutional Networks. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, volume 8689, 818–833. Cham: Springer International Publishing. ISBN 978-3-319-10589-5 978-3-319-10590-1.

Zheng, L.; Guha, N.; Anderson, B. R.; Henderson, P.; and Ho, D. E. 2021. When Does Pretraining Help?: Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 159–168. São Paulo Brazil: ACM. ISBN 978-1-4503-8526-8.