Do Charge Prediction Models Learn Legal Theory?

Zhenwei An[1,2]*, **Quzhe Huang**[1,3]*, **Cong Jiang**[4,5],
✉**Yansong Feng**[1,6] and **Dongyan Zhao**[1,5,6]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]School of Software & Microelectronics, Peking University
[3]School of Intelligence Science and Technology, Peking University
[4] Peking University Law School [5] Institute for Artificial Intelligence, Peking University
[6] The MOE Key Laboratory of Computational Linguistics, Peking University
{anzhenwei,huangquzhe,jiangcong,fengyansong,zhaody}@pku.edu.cn

Abstract

The charge prediction task aims to predict the charge for a case given its fact description. Recent models have already achieved impressive accuracy in this task, however, little is understood about the mechanisms they use to perform the judgment. For practical applications, a charge prediction model should conform to the certain legal theory in civil law countries, as under the framework of civil law, all cases are judged according to certain local legal theories. In China, for example, nearly all criminal judges make decisions based on the Four Elements Theory (FET). In this paper, we argue that trustworthy charge prediction models should take legal theories into consideration, and standing on prior studies in model interpretation, we propose three principles for trustworthy models should follow in this task, which are `sensitive`, `selective`, and `presumption of innocence`. We further design a new framework to evaluate whether existing charge prediction models learn legal theories. Our findings indicate that, while existing charge prediction models meet the `selective` principle on a benchmark dataset, most of them are still not `sensitive` enough and do not satisfy the `presumption of innocence`. Our code and dataset are released at https://github.com/ZhenweiAn/EXP_LJP.

1 Introduction

The task of charge prediction is to determine appropriate charges, such as *Fraud* or *Theft*, for a case by analyzing its textual fact descriptions. Such a technique is beneficial for improving the efficiency of legal professionals, e.g., helping judges, lawyers, or prosecutors to distinguish similar charges and focus on discriminative features. But as an auxiliary tool in the legal domain, it should be used with great caution, in case of introducing undesirable unfairness (Angwin et al., 2016).

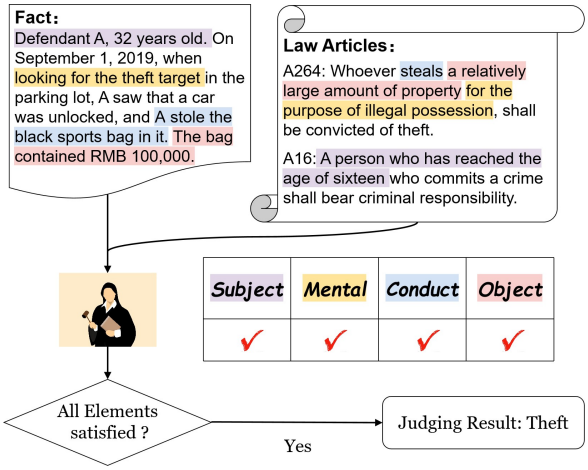

Figure 1: An example of accusing the defendant of *Theft*. FET is the most dominant legal theory in China, which defines that a case must satisfy four criminal elements simultaneously to constitute a crime

Most existing works formalize charge prediction as a text classification task (Luo et al., 2017; Hu et al., 2018; Zhong et al., 2018). Although recent advances in deep learning have demonstrated their excellent performance in predicting the charges (Yang et al., 2019; Xiao et al., 2021), their reliability and interpretability are still underexplored. It is unknown whether the intrinsic decision mechanism of these models corresponds to the decision logic of human judges. Specifically, since most existing models are data-driven and all cases in the charge prediction dataset conform to local legal theories, it is necessary to figure out whether these charge prediction models learn their corresponding legal theories.

Previous studies have shown that trustworthy legal AI models are supposed to point out human-interpretable factors used in a decision (Atkinson et al., 2020). Besides, they should also explain how the changes in fact descriptions would change their decisions. Based on these discussions, we argue that a trustworthy charge prediction model should

* Equal Contribution.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3757–3768
December 7-11, 2022

obey the following principles to conform to local legal theory and illustrate how they act in legal perspectives using FET, the most dominant legal theory in China (Wang, 2017), as an example:

1) **`Selective`**: be able to identify and concentrate on important parts of a case when making decisions. In FET, the important parts are considered as *criminal elements*. 2) **`Sensitive`**: be aware of the subtle distinctions between similar charges. When three of the four criminal elements in FET are identical for a pair of similar charges, a trustworthy model is expected to use the remaining criminal element to distinguish the similar charges.

Apart from the prerequisites, which have been extensively explored in various domains, we can not ignore the presumption of innocence when focusing on a legal task. Presumption of innocence refers to the principle that any defendant is presumed innocent until proven guilty in a criminal trial, which is fundamental to protect human rights worldwide (Tadros and Tierney, 2004). Taking this presumption into account, we propose an additional principle that a trustworthy charge prediction model should follow: 3) **`Presumption of innocence`**: always assume innocent unless sufficient requirements for a charge are met. In FET, `presumption of innocence` is guaranteed by checking all four criminal elements before making decisions.

In this paper, we propose a framework to evaluate whether a charge prediction model conforms to certain legal theory. Our framework consists of three components that evaluate the aforementioned principles respectively. We first apply a probing task to measure whether models learn the skill of identifying criminal elements from fact descriptions, corresponding to the `selective` principle. The assumption here is that if the model is capable of identifying criminal elements, the knowledge of such a skill should be reflected in its internal representations, which could be detected by a diagnostic model (Alt et al., 2020).

The evaluation of the `sensitive` principle relies on a perturbation experiment, in which we modify the fact descriptions of confusing charges and check whether the model could detect the modifications. Specifically, for a pair of confusing charges, we rewrite the fact descriptions related to a certain criminal element and make the modified facts fulfill the requirements of the other charge. If a model is `sensitive` enough, it should be capable of identifying these modifications and making different predictions for the original facts and the modified ones. The final component evaluates whether models follow the `presumption of innocence` by checking the model's performance on incomplete fact descriptions. Those incomplete facts are obtained by excluding all descriptions related to a specific criminal element from criminal descriptions. The models are expected to make innocent predictions for those incomplete fact descriptions, because they violate the requirements of FET that all the four criminal elements should be satisfied when judging guilty.

We conduct experiments with popular Chinese charge prediction models and the results indicate that, while existing charge prediction models meet the `selective` principle on our benchmark dataset, most of them are still not `sensitive` enough and do not satisfy the `presumption of innocence`.

Our contributions are four-folds: (1) We propose the first ever set of principles that a trustworthy charge prediction model should follow when conforming to certain legal theories. (2) Based on these principles, we propose a new investigation framework to evaluate the trustworthiness of charge prediction models. (3) We supplement the current popular charge prediction dataset CAIL (Xiao et al., 2018) with innocent cases and provide sentence-level criminal elements annotation for a subset. (4) We examine existing Chinese charge prediction models using FET, the most widely used legal theory in China, on the new benchmark, and find that most existing charge prediction models are not trustworthy enough, though they can achieve over 80% prediction accuracy.

2 The Charge Prediction Task

Suppose the fact description of a case is a word sequence $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$, where n is the length of $\mathbf{x}$. Based on the fact description $\mathbf{x}$, the charge prediction task aims at predicting an appropriate charge $y \in Y$, where Y is the potential charge set.

To solve this task, previous works often use existing text classification models (Li et al., 2018; RN5), many of which are later improved by introducing legal knowledge (Luo et al., 2017; Zhong et al., 2018; Yang et al., 2019). More recently, pretrained language models have also been proven effective in this task (Xiao et al., 2021).

In our study, we select the following representa-

tive charge prediction models to evaluate whether they are trustworthy according to the specific legal theory, i.e., the FET in this case.

BiLSTM Luo et al. (2017) uses Bi-LSTM (Yang et al., 2016) to encode fact descriptions and applies an attention mechanism to aggregate encoded word representations to obtain fact embedding, which is then used for classification.

TopJudge TopJudge (Zhong et al., 2018) is a representative of those multitask learning models. During encoding, TopJudge employs CNN (Kim, 2014) as the encoder to obtain fact embeddings. In decoding, it exploits a directed acyclic graph to capture the relationship among three sub-tasks, i.e., charge prediction, law article prediction, and term prediction, which are jointly optimized in a multitask framework.

FewShot FewShot (Hu et al., 2018) introduces discriminative attributes to distinguish confusing charges and provide additional knowledge for few-shot charges, which can stand for those models that introduce legal knowledge into the charge prediction task. It uses LSTM (Hochreiter and Schmidhuber, 1997) as the fact encoder and conducts charge prediction and attributes prediction afterward.

BERT BERT (Devlin et al., 2019) is a strong baseline for many text classification tasks. We use the representation of [CLS] token for classification.

Lawformer Xiao et al. (2021) is a Longformer-based (Beltagy et al., 2020) language model, which is pretrained on large-scale Chinese legal cases. We use it to encode the fact description and apply the classification based on the [CLS] token.

2.1 The Four Elements Theory

Legal theories are the bases for judges to correctly determine charges, which define the method of analyzing cases. Judges are required to follow legal theories when making judgements (Gao, 1993). If they do not, they might make decisions arbitrarily, which is a breach of human rights and freedom (Wang, 2017).

In China, the Four Elements Theory (FET) is the dominant legal theory for criminal trials. In practice, nearly all criminal judges use FET to justify their decisions (Jiyao, 2011). As a result, a trustworthy Chinese charge prediction model should also conform to FET since they are trained based on the judgment documents which conform to the local legal theory, FET.

	Acc	F1	P	R
TopJudge	82.7	60.6	67.5	59.2
FewShot	82.9	71.7	75.9	71.6
BiLSTM	82.4	59.8	65.7	58.9
Bert	90.4	81.9	83.2	79.8
Lawformer	91.0	83.8	84.4	81.1

Table 1: Charge Prediction results on CAIL-I, where Acc, F1, P, and R represent Accuracy, macro F1, macro precision, and macro recall, respectively.

According to FET, a case must satisfy four criminal elements simultaneously to constitute a crime. The four criminal elements are: (1) the *subject (Sub)* refers to the person or organization who has committed the criminal offense and shall bear criminal crimes, (2) the *object (Obj)* refers to the person, thing, interest, or social relations protected by criminal law and jeopardised by criminal offence, (3) the *conduct (Con)* refers to harmful behaviors, and (4) the *mental state (Men)* is the mental state of the criminal subject when committing a crime, either *intent* or *negligence*.

For example, the four criminal elements of *Theft* are as follows: (1) *subject*: the general subject, that is, a person who has reached the age of criminal responsibility (16 years old in China), (2) *object*: public or private property, (3) *conduct*: the act of stealing a large amount of property or repeatedly stealing property, (4) *mental state*: intent and with the purpose of illegal possession.

3 Dataset

Existing charge prediction datasets, such as CAIL (Xiao et al., 2018), have played a crucial role in the development of legal artificial intelligence research. However, they suffer from two limitations: (1) Lacking innocent cases. This violates the presumption of innocence, one of the most fundamental legal principles worldwide. (2) Only containing coarse-grained annotations, such as charges and law articles, which cannot reveal how the judges analyze the cases.

To alleviate the two shortcomings, in this paper, we propose a new charge prediction dataset, CAIL-I, that adds innocent cases to the original CAIL. We further annotate whether a sentence is related to certain criminal elements in a subset of CAIL. We call this Sentence-level Criminal Elements dataset as SCE, which can be utilized to analyze whether a

Charge	Sub	Men	Con	Obj	NA	Cases
TA	102	228	258	163	833	100
Rob	109	185	435	153	673	98
FS	125	173	281	187	678	99
Cor	120	132	361	176	595	94
MoF	122	118	289	127	524	100
MoPF	122	133	318	135	571	97
NH	105	192	312	169	711	97
All	805	1161	2254	1110	4585	685

Table 2: Statistics of Sentence-level Criminal Elements dataset (SCE). Columns 2-5 show the number of sentences involving different criminal elements, where NA means being related to none criminal elements. The last column indicates the number of cases corresponding to different charges. The abbreviations of criminal elements and charges are clarified in Section 3.

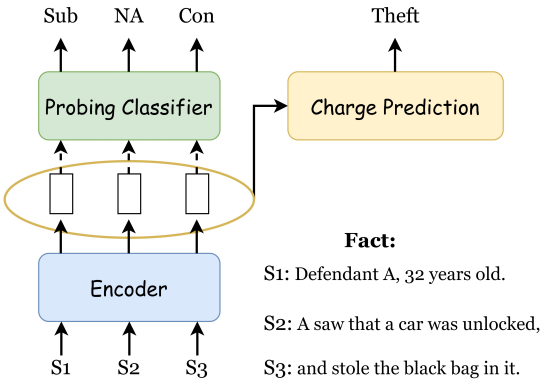

Figure 2: Probing Setup. The dotted arrows emphasize that probing is applied to frozen encoders after training of charge prediction. Only the parameters of the linear classifier module are learned during probing.

model conforms to FET.

Collecting Innocent Cases To obtain innocent cases, we first collect all non-prosecution cases from the Chinese Prosecutor's Website[1]. Among these non-prosecution cases, the real innocent cases take only small part. Many cases are not prosecuted for other reasons, such as the defendant died before prosecution. To identify the real innocent cases, we hire 2 law school graduate students to review the collected data case by case, and only kept the cases with truly innocent defendants. Finally, we obtain 462 innocent cases and add them into the CAIL training set, validation set, and test set at the ratio of 5:3:2 to form the new benchmark CAIL-I. We report the performance of existing charge prediction models on CAIL-I in Table 1.

Annotating Criminal Elements Given a fact description and the corresponding charge label, annotators are asked to label each sentence with related criminal elements or NA when the sentence does not relate to any criminal elements. As a sentence might contain information about various criminal elements, it could be annotated with more than one label. This annotation needs substantial legal knowledge involvement and the fine-grained scheme requires a huge workload, thus, it is impossible to annotate the whole CAIL-I datasets. To alleviate the burden of manual annotation, we choose 7 charges which are hard to distinguish in practice (Ouyang, 1999). These confusing charges could help us better understand models' behavior under FET. The 7 charges are *Traffic accident (TA), Robbery (Rob), Forcible seizure (FS), Corruption (Cor), Misappropriation of funds (MoF), Misappropriation of public funds (MoPF)* and *Negligent homicide (NH)*. We employed 2 paid graduate students from law schools as annotators and the inter-annotator Cohen's Kappa is 0.64. Table 2 shows the statistics of this Sentence-level Criminal Elements dataset (SCE).

[1]www.12309.gov.cn

4 Selective Principle Checking

Judges are required to extract criminal elements and filter less important information from fact descriptions, where the first step is to relate each sentence to its corresponding criminal elements. However, it is unclear whether existing charge prediction models are `selective` enough to relate sentences to criminal elements. To figure it out, we design a probing task to explore these models' ability to distinguish criminal elements of a charge.

Probing is a popular approach to model introspection, which trains a simple classifier – a probe, to predict certain desired information from the latent representations learned by neural networks. High prediction performance is interpreted as evidence for the information being encoded in the representations and indicates that the information is what the neural networks rely on (Saleh et al., 2020; Alt et al., 2020).

Our probing task is to examine whether the representations have encoded the type of criminal elements, which represents the model's ability to identify them, i.e., the `selectivity` of models. Figure 2 shows our probing setup. Specifically, we freeze encoders of charge prediction models and obtain the sentence representations of fact descrip-

Models	TA	NH	FS	Rob	MoF	MoPF	Cor
Random	14.3/25.3/19.3	18.1/28.9/22.3	16.3/23.6/19.5	22.3/33.4/26.9	18.7/28.5/22/6	20.2/30.8/24.4	21.1/32.1/25.5
TopJudge	83.5/64.7/72.9	75.0/48.7/59.0	69.8/40.1/50.9	70.5/45.1/55.0	78.5/58.6/67.1	72.5/47.4/57.3	73.9/49.7/59.5
FewShot	80.4/82.8/81.6	67.6/65.7/66.6	64.3/57.6/60.7	59.5/52.9/56.0	72.2/65.4/68.6	64.1/52.5/57.7	63.1/53.4/57.8
BiLSTM	86.6/82.8/84.7	72.3/65.6/68.8	68.8/54.4/60.8	69.6/53.3/60.4	80.8/62.3/70.4	72.1/51.2/59.9	71.9/55.8/62.8
BERT	83.1/86.7/84.9	71.5/66.8/69.1	66.8/53.4/54.5	66.2/54.9/60.0	78.2/63.4/70.0	72.6/54.3/62.1	68.4/56.7/62.0
Lawformer	84.5/83.2/83.8	72.8/66.6/69.5	66.7/55.5/60.6	67.2/55.7/60.9	80.8/66.6/73.0	70.9/52.1/60.0	73.1/60.1/66.0
ELMO*	85.6/84.4/84.9	75.2/67.1/70.9	70.9/55.5/62.2	67.5/53.7/59.9	81.0/71.0/75.7	72.1/57.7/64.1	73.2/62.2/67.3
BERT*	85.8/85.8/85.8	74.4/66.3/70.1	73.2/56.3/63.6	71.9/57.4/63.8	80.9/69.2/74.6	79.3/57.2/66.5	78.1/58.7/67.0
Lawformer*	85.6/86.3/85.9	76.2/65.2/70.3	70.5/55.4/62.0	69.6/58.5/63.6	82.9/70.0/75.9	78.2/56.3/65.5	76.8/59.2/66.9

Table 3: Precision/Recall/F1 of probing results for every charge in SCE. We report the average micro-metrics (%) over 5 folds. The baseline performance is reported at the top and the performances of language models are shown at the bottom, indicated by *.

tions by the encoders. Then a linear classifier is trained on these representations to predict whether a sentence is related to certain criminal elements. This follows the same methodology used in previous works (Saleh et al., 2020; Alt et al., 2020).

We apply mean pooling to get the sentence representation from the word embeddings encoded by the encoder of specific charge prediction models. In order to explore how the charge prediction task influences models' ability to identify criminal elements, we conduct probing experiments on a few language models, including ELMO (Peters et al., 2018), BERT, and Lawformer. The experiment is conducted on the SCE dataset.

Table 3 shows the result of probing. We do 5-Fold Cross-Validation on SCE and report the average. We use Random as a baseline, where we randomly assign the label to every sentence, based on the frequency of each label in the training set.

Capacity of being `selective` As shown in Table 3, all the charge prediction models outperform the baseline Random substantially. The good performance of charge prediction models in the probing task indicates that those models have learned the skill of identifying criminal elements and has the ability to distinguish them. In other words, existing charge prediction models are capable of being `selective`.

Effect of semantic information It is surprising to find that the language models, which are not finetuned on charge prediction, also perform well in the probing task. For example, BERT* achieves over 60% micro-F1 scores for all the charges. This is because semantic information is enough for identifying criminal elements in many circumstances. Taking the phrase "*car crush*" as an example, it is easy to connect it with the *conduct* element of *Traffic accident*, when understanding this phrase describes a car hitting something. Even without legal knowledge, one will not consider "car crush" as introducing who was involved in an accident, i.e., the *subject* element. Instead of the involvement of legal knowledge, understanding such phrases requires comprehending the semantic information of words or phrases, which has been successfully captured by language models like BERT.

Bias from shortcut Another interesting finding is that models that have been trained for charge prediction perform worse in the probing task, e.g., BERT performs worse than BERT*. By comparing the predictions of these two groups of models, we discover that the performance drop in probing is due to the bias for particular patterns learnt by charge prediction models . During training, models learn that some patterns are highly correlated with specific charges, providing a shortcut for making judgments. Such a high correlation encourages the models to believe that those patterns are associated with specific criminal elements, a bias that ultimately results in the charge prediction models' worse performance in the probing task. This bias will lead to an incorrect association of sentences with certain criminal elements, which is evidenced by the fact that the decline of F1 is largely due to precision rather than recall, as shown in Table 3.

A took advantage of the job convenience to keep 131,000 yuan of the company's business case, which A had not paid back yet.

[Conduct Element]

The above case shows a detailed example of such bias. When learning charge prediction, Lawformer recognizes that the pattern of *did not pay back money* is highly correlated with the *conduct*

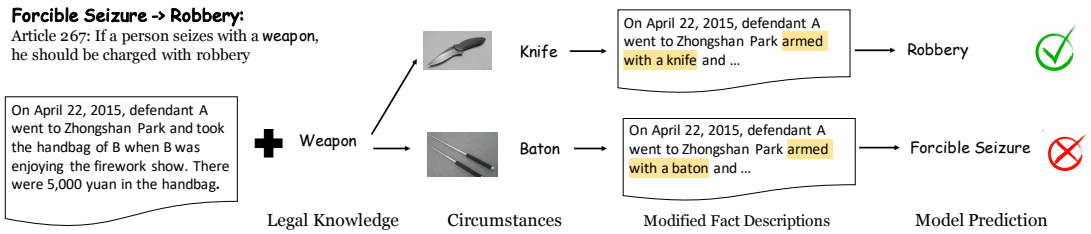

Figure 3: An example for sensitive principle checking.

element of *Misappropriation of funds* without considering its context.

According to Chinese criminal law, the *conduct* element of *Misappropriation of funds* corresponds to the event where the defendants plunder money for more than three months and do not give the plundered money back before they are investigated. In most cases, the text pattern *did not pay back money* implies the *conduct* element of *Misappropriation of funds*. However, in the case indicated by the following example, it does not imply any criminal element.

> B misappropriated 140,000 yuan for personal costs.
> After being investigated , B did not paid back the money
> [Not Element]

In this case, *after being investigated* indicates that the crime had been completed and the legal authority began to investigate the crime. The act *did not pay back money* that happened after investigation cannot be used for charge prediction, hence does not belong to the *conduct* element of *Misappropriation of funds*.

5 Sensitive Principle Checking

A reliable charge prediction model should be sensitive to the subtle difference between the fact descriptions of confusing charges. In this section, we collect pairs of similar fact descriptions and evaluate whether existing models could recognize the difference.

Specifically, we select three groups of confusing charges where the charges in each group differ in only one criminal element from each other. According to the difference between the criminal elements of two charges, we modify the fact descriptions and make the modified ones meet the requirements of the other charge in the same group. We expect that a reliable model could recognize the distinctions and make different predictions for the original fact descriptions and the modified ones.

As shown in Figure 3, the difference between *Forcible Seizure* and *Robbery* lies in their *conduct* elements. "armed with a weapon or not" is the representative legal knowledge in distinguishing *Robbery* from *Forcible seizure*, as "armed with a weapon" could bring coercion to the victim, which is the *conduct* element of *Robbery*. Then, we add descriptions "armed with a knife" and "armed with a baton" to the fact descriptions of *Forcible Seizure*. A trustworthy charge prediction model is expected to learn such legal knowledge during training and predict *Robbery* for the modified fact descriptions.

For legal knowledge, like "armed with weapons" in Figure 3, we design two specific circumstances, e.g., "armed with a knife" and "armed with a baton", where "armed with a knife" is much more common than the other. This design is to determine if the charge prediction models truly learn the legal knowledge, rather than simply remembering common textual patterns. If the model only recognizes the common circumstance and ignores the other, we believe the model does not learn that legal knowledge.

Changes	Legal Knowledge	Specific Circumstances
FS→Rob	armed with weapon	armed with a baton
		armed with a knife
TFT→Rob	using violence	spray the security guards with pepper
		hurt pursuers with a switchblade
TA→NH	on non-public transport road	on a road where the sewer is being repaired
		on a road closed for construction

Table 4: The legal knowledge and their corresponding circumstances used to modify the fact descriptions.

Table 4 lists the three pairs of confusing charges and corresponding legal knowledge. For each pair, we randomly select 200 cases from the validation

Charge	FS → Rob		TFT → Rob		TA → NH	
Circumstance	△	*	△	*	△	*
TopJudge	73.5	45.5	63.5	36.0	87.5	89.5
FewShot	93.5	91.0	87.5	83.5	87.5	84.0
BiLSTM	82.5	83.0	77.5	47.0	92.0	89.5
BERT	88.0	14.5	88.0	26.0	96.0	79.5
Lawformer	84.0	48.0	59.5	33.5	98.0	97.5

Table 5: The ratio of predicting the original charges after perturbations. The "△" refers to the more uncommon circumstance and "*" refers to the common one.

	Subject	Mental	Conduct	Object
BiLSTM	0.844	0.777	0.572	0.693
TopJudge	0.772	0.666	0.574	0.644
FewShot	0.826	0.753	0.619	0.740
BERT	0.920	0.866	0.704	0.826
Lawformer	0.924	0.880	0.712	0.841

Table 6: The consistency of models' predictions between using the complete descriptions and the modified descriptions after removing the expressions related to one criminal element.

and test sets of CAIL-I and modify those fact descriptions with the two specific circumstances. The results are summarized in Table 5.

Able to distinguish confusing charges? In most cases, charge prediction models still predict the original charge when the modified fact descriptions no longer satisfy the original one. Among three pairs of confusing charges, models perform best in distinguishing the *conduct* element between *Theft* and *Robbery*, although the ratio of predicting the original charge still exceeds 50%. When it comes to *Traffic accident* and *Negligent homicide*, this ratio even reaches around 100% for some models, indicating that these models totally fail to recognize the difference. It is surprising that FewShot does not perform well in this task, as it requires the model to pay explicit attention to several legal attributes. We think this is because their attributes are too coarse-grained and sparse. FewShot only designs 10 legal attributes for over 200 charges, and some legal knowledge, like "on non-public transport road", are not considered. The highly abstracted attributes may be useful in few-shot settings, but they cannot make the model more `sensitive`. Overall, the poor performance in distinguishing confusing charges indicates that the existing models are not `sensitive` enough.

Textual patterns or legal knowledge? There are obvious discrepancies between models' capacities to recognize two distinct circumstances of the same legal knowledge. Taking BERT as an example, it cannot identify the uncommon circumstance "armed with a baton" for 88% cases in the setting of *Forcible seizure* and *Robbery*. However, the models are very sensitive to the common one, "armed with a knife", with only 14.5% cases maintaining the original prediction. The distinct performance suggests that charge prediction models are more likely to remember common textual patterns instead of understanding the legal knowledge necessary to discriminate between criminal elements of confusing charges.

6 Presumption of innocence Checking

In China, FET guarantees the presumption of innocence by checking the completeness of all four criminal elements. The theory requires that only when all four criminal elements are satisfied, will a defendant be convicted of that charge. Based on this completeness checking, although A in the following example satisfied three criminal elements, A was innocent because A did not intend to occupy the phone and did not fulfill the requirement of the *mental state* element.

> A sat next to B on the bus. The wallet of B slipped out of B's pocket just before B got off the bus. A picked it up and got off as far as possible to give it back to B , but A did not find B. When B realize the wallet was lost, B called the police. Then the police arrested A on suspicious of the theft.

Ideally, a reliable model, which follows the `presumption of innocence`, should have the ability to check whether all the four criminal elements are satisfied and predict a case as innocent when its fact description lacks one or more criminal elements. In this section, we explore whether existing charge prediction models have such an ability. Specifically, we generate cases lacking one criminal element by removing all the sentences related to that element in criminal fact descriptions and check whether the model could change its predictions when identifying such modifications. We use the SCE dataset to generate attack cases.

The ratio of predicting the same charge after removing one criminal element is shown in Table 6, and the confidence densities for predicting the original charge using the complete descriptions and the descriptions after removing the element-specific

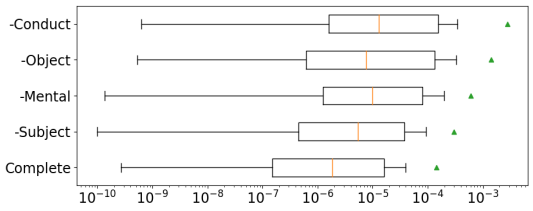

Figure 4: Boxplot describing Lawformer's prediction probability of innocent. "Complete" means complete fact descriptions. "-Subject" means fact descriptions with all sentences relevant to *subject* element removed. So do "-Conduct", "-Object" and "-Mental".

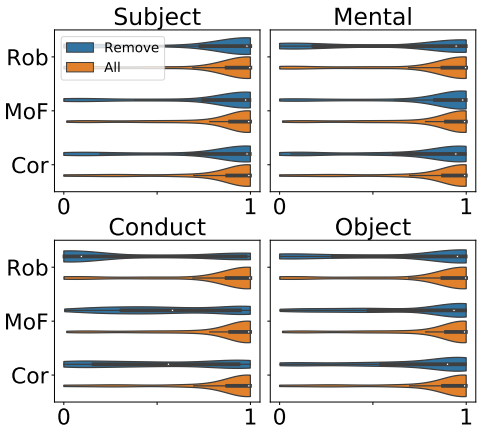

Figure 5: Confidence densities of predicting the original charges using the complete fact (orange) and using the fact after removing the descriptions related to a specific criminal element (blue).

expressions are shown in Figure 5. The results of all charges and models are shown in Appendix. We can draw the following conclusions from Table 6 and Figure 5:

Charge prediction models do not satisfy the `presumption of innocence`. While charge prediction models are expected to recognize the absence of any criminal element, as shown in Table 6, all models stick to their predictions with incomplete fact descriptions most of the time. Taking the *subject* element as an example, when the descriptions related to it are deleted, TopJudge, the best-performing model, can maintain its predictions with a ratio of about 80%, and for BERT and Lawformer, this ratio even exceeds 90%. But it does not mean that these models completely ignore the absence of some elements. We discover that when a criminal element is removed, models improve the prediction probability of innocent, as shown in Figure 4. However, the improvement is insufficient to satisfy the `presumption of innocence`.

Which criminal elements gain more attention? As illustrated in Figure 5, Lawformer's confidence density changes significantly when the *conduct* element is omitted. For the remaining three criminal elements, eliminating the relevant fact descriptions has little effect on the Lawformer's confidence in predicting the original charge, and in most cases, the confidence remains greater than 80%, which is a very high level. The results of other models are shown in the Appendix and the results are similar as Lawformer's.

7 Related Work

Probing Probing is a popular method for model introspection, which associates the representations learned by the neural networks with properties of interests and examines the extent to which these properties can be recovered from the representations (Ravichander et al., 2021; Adi et al., 2017). Previous works mainly focus on what linguistic properties models have learnt (Belinkov et al., 2017; Tenney et al., 2019; Warstadt et al., 2019; Vulić et al., 2020). There also has been research that employs probing to investigate the mechanisms that models used to perform certain tasks(Alt et al., 2020; Saleh et al., 2020). Although the probing method is widely used, as far as we know, no research has employed probing to explore whether neural networks could learn legal knowledge, like the Four Elements Theory in our paper. More discussion of probing could be found in (Belinkov and Glass, 2019; Belinkov, 2022)

Besides, we are not the first to analyze models by manipulating the input text. Many studies point out that minor changes in the text may bring unexpected results from the neural models, like Bert-Attack (Li et al., 2020). But little work has been done in the field of analyzing legal texts.

Interpretable Charge Prediction Models When machine learning algorithms are put into practice for automated individual decision making, many legal experts demand *right to explanation* (Doshi-Velez et al., 2017) for these algorithms, whereby users especially the losing party in a justification have the right to ask for an explanation of an algorithm decision that significantly affects them. Consequently, the reliability and interpretability of charge prediction models are of equal importance to their performance.

To improve the interpretability of charge predictions, several studies generate charge prediction alongside its supports. Jiang et al. (2018) and Liu et al. (2018) employ reinforce learning to derive rationales at the phrase level to explain the model's output. Hu et al. (2018) and Li et al. (2019) design certain attributes to help distinguish confusing charges. They train a classifier for each attribution and then make decisions depending on whether or not the fact descriptions satisfy those attributes. Zhong et al. (2020) designs a series of questions and solves charge prediction by answering those questions, with each question corresponding to a certain attribute. Liu et al. (2021) exhibits important evidence for judgment with causal graphs and causal chains. Li2 achieves multi-granularity inference of legal charges by obtaining the subjective and objective elements from the fact descriptions of legal cases.

Although these strategies are more interpretable than those that just provide a charge, their reasoning procedure may still violates FET. Some of them, such as Li et al. (2019), are solely concerned with the life-related *object* element and disregard money-related *object* element, thus cannot perform criminal element extraction for all charges. Some other works concentrate exclusively on a subset of the four criminal elements, omitting the others. Hu et al. (2018), Zhong et al. (2020), Liu et al. (2021), and Li2 do not consider the subject element, implying that they do not check whether all the elements are satisfied when making decisions, thus violate the `presumption of innocence`. As a result, the reliability of those models remains questionable.

8 Conclusion and Future Work

When applying artificial intelligence in the domain of law, not only do we expect a model to achieve high accuracy, but we also require the model to be trustworthy. Our work proposes three principles that a trustworthy model should follow in the charge prediction task based on both previous efforts on explanation and legal theories. According to the principles, we examine existing charge prediction models and our analysis shows that while they satisfy the `selective principle`, most models are not `sensitive` enough and do not satisfy the `presumption of innocence`. We hope our discoveries will help the Artificial Intelligence and Law community better understand the mechanism of charge prediction models. We suggest the following directions for future work:

- Extend current datasets with innocent cases to ensure models trained on them satisfy the `presumption of innocence`.
- Help models understand legal knowledge instead of identifying certain patterns.
- Design models which can perform completeness checking of all criminal elements before convicting the defendant of guilty.

Limitations

Although this paper proposes principles of reliable charge prediction models based on legal theory and develops a framework for determining if a charge prediction model learns certain legal theory, we do not present a model which can actually adhere to these principles. It requires more exploration and research from the AI and Law community. In addition, the experiment designed to examine the `sensitive` principle requires substantial annotations from legal experts, making it inconvenient in extending such method to other legal theories. Lastly, due to limited public criminal cases, we are only able to collect a subset of innocent instances.

Ethical Consideration

Intended Use Our work could help the community of AI and Law better understand the mechanism of existing charge prediction models. We illustrate that the existing charge prediction models do not conform to the legal theory in China, and we call for using these models with more caution.

Misuse Potential Our work shows that existing charge prediction models are `selective` in our dataset, but that does not mean those models conform to the Four Element Theory. We think the existing charge prediction models could not replace judges and make predictions independently.

Acknowledgements

This work is supported in part by NSFC (62161160339) and National Key R&D Program of China (No. 2018YFC0831900). We would like to thank the anonymous reviewers for their helpful comments and suggestions; thank Chen Zhang, Xiao Liu and Weiye Chen for providing feedback on an early draft. For any correspondence, please contact Yansong Feng.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Probing linguistic features of sentence-level representations in neural relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1534–1545, Online.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications.

Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. 2020. Explanation in ai and law: Past, present and future. *Artificial Intelligence*, 289:103387.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

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

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.

Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI under the law: The role of explanation. *ArXiv preprint*, abs/1711.01134.

Mingxuan Gao. 1993. *The theory of criminal jurisprudence*. China Renmin University Publisher.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

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.

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.

Tang Jiyao. 2011. Criminal theory system in germany japan and the system of constitution of crime:a dissection,reflection and reference under empirical study. *Science of Law(Journal of Northwest University of Political Science and Law)*, 29:68–81.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online.

Penghua Li, Fen Zhao, Yuanyuan Li, and Ziqin Zhu. 2018. Law text classification using semi-supervised convolutional neural networks. In *2018 Chinese Control And Decision Conference (CCDC)*.

Shang Li, Boyang Liu, Lin Ye, Hongli Zhang, and Binxing Fang. 2019. Element-aware legal judgment prediction for criminal cases with confusing charges. In *31st IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2019, Portland, OR, USA, November 4-6, 2019*, pages 660–667. IEEE.

Xianggen Liu, Lili Mou, Haotian Cui, Zhengdong Lu, and Sen Song. 2018. Jumper: Learning when to make classification decision in reading. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4237–4243. ijcai.org.

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. 2021. Everything has a cause: Leveraging causal inference in legal text analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1928–1941, Online.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736, Copenhagen, Denmark.

Tao Ouyang. 1999. *Confusing crimes, noncrime, and boundaries between crimes*, volume 1.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online.

Abdelrhman Saleh, Tovly Deutsch, Stephen Casper, Yonatan Belinkov, and Stuart Shieber. 2020. Probing neural dialog models for conversational understanding. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 132–143, Online.

Victor Tadros and Stephen Tierney. 2004. The presumption of innocence and the human rights act. *The Modern Law Review*, 67(3):402–434.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online.

S. Wang. 2017. *Criminal law in china*. Wolters Kluwer.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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 2877–2887, Hong Kong, China.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*.

Chaojun Xiao, Haoxiang Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *ArXiv*.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4085–4091. ijcai.org.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1480–1489. The Association for Computational Linguistics.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively questioning and answering for interpretable legal judgment prediction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 1250–1257. AAAI Press.

Appendix

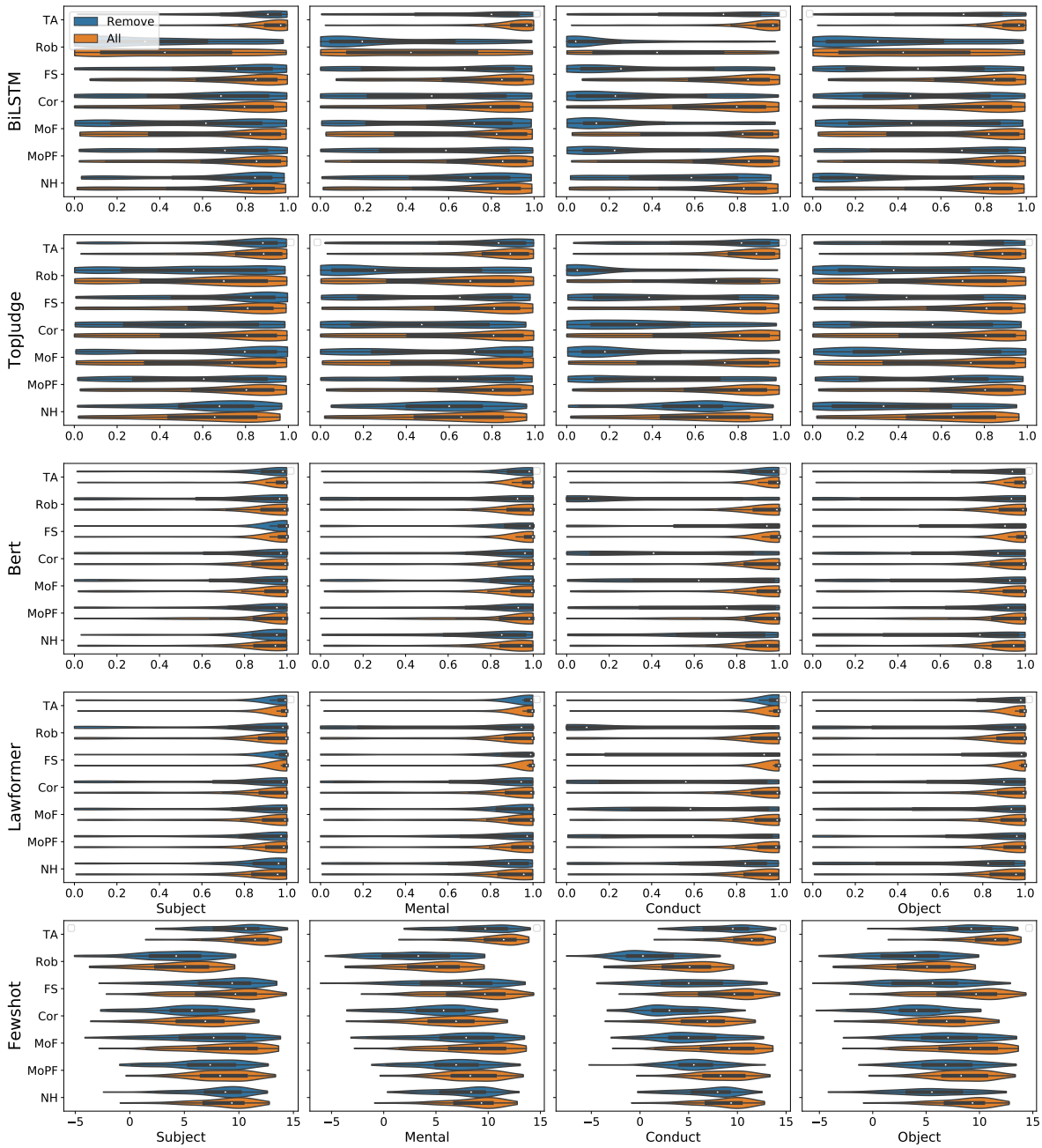

Figure 6: Confidence densities of predicting the original crimes using the complete fact (orange) and using the fact after removing the descriptions related to a specific criminal element (blue).