# ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US

**Gil Semo*** **Dor Bernsohn** **Ben Hagag** **Gila Hayat** **Joel Niklaus***†
Darrow AI Ltd. Niklaus.ai
30 Ha'arbaa Street, Tel Aviv, Israel Schwarztorstrasse 108, Bern, Switzerland
firstname.lastname@darrow.ai joel@niklaus.ai

## Abstract

The research field of Legal Natural Language Processing (NLP) has been very active recently, with Legal Judgment Prediction (LJP) becoming one of the most extensively studied tasks. To date, most publicly released LJP datasets originate from countries with civil law. In this work, we release, for the first time, a challenging LJP dataset focused on class action cases in the US. It is the first dataset in the common law system that focuses on the harder and more realistic task involving the complaints as input instead of the often used facts summary written by the court. Additionally, we study the difficulty of the task by collecting expert human predictions, showing that even human experts can only reach 53% accuracy on this dataset. Our Longformer model clearly outperforms the human baseline (63%), despite only considering the first 2,048 tokens. Furthermore, we perform a detailed error analysis and find that the Longformer model is significantly better calibrated than the human experts. Finally, we publicly release the dataset and the code used for the experiments.

## 1 Introduction

Recently, the literature in Legal Natural Language Processing (NLP) has grown at a fast pace, firmly establishing it as an important specialized domain in the broader NLP ecosystem. As part of this strong growth and as a first step establishing Legal NLP in the field, many legal datasets have been released in the fields of Legal Judgment Prediction (LJP) (Niklaus et al., 2021a; Chalkidis et al., 2019), Law Area Prediction (Glaser and Matthes, 2020), Legal Information Retrieval (Wrzalik and Krechel, 2021), Argument Mining (Urchs et al., 2022), Topic Classification (Chalkidis et al., 2021a), Named Entity Recognition (Luz de Araujo et al., 2018; Angelidis et al., 2018; Leitner et al., 2019), Natural Language Inference (Koreeda and Manning, 2021), Question Answering (Zheng et al., 2021; Hendrycks et al., 2021), and Summarization (Shen et al., 2022; Kornilova and Eidelman, 2019).

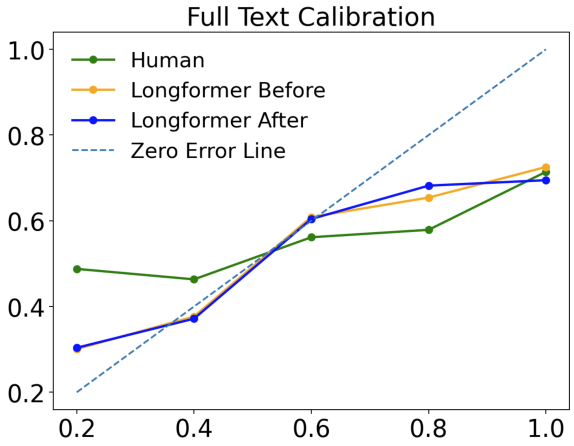


Figure 1: Calibration plot on the Full Text dataset. The human experts rated the confidence of their predictions on a score from 1 to 5. The confidence scores of the Longformer models were binned into 5 buckets.

In particular, the field of LJP has been very active, with many datasets released recently. Cui et al. (2022) surveyed the field and divided the datasets into five subtasks. In this work, we release a dataset belonging to the category of the Plea Judgment Prediction (PJP) task. Most other PJP datasets use the facts summary, written by the court (clerks or judges) as input (Cui et al., 2022). The facts are written in such a way as to support the final decision (Niklaus et al., 2021a) and require extensive work by highly qualified legal experts (Ma et al., 2021). In contrast, in this work we consider the plaintiff's pleas (AKA complaints) as input, making the task more realistic for use in real-world applications.

Most LJP datasets released so far are from countries with civil law. Our dataset originates from the United States, the largest country employing the common law legal system. To the best of our knowledge, we are the first to release a dataset specifically targeting class action lawsuits.

* Equal Contribution
† Corresponding Author



*Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 31 - 46

**Motivation**

The 16th United Nations Sustainable Development Goal (UNSDG) is to "Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels". Class actions are a private enforcement instrument that enables courts to organize the mass adjudication of meritorious claims by underrepresented individuals and communities. Without class actions, many victims of illegal action would never get their day in court. Making case outcomes and facts accessible is crucial to strengthen the effective use of class actions and private enforcement to drive UNSDG 16. With the power of early LJP, plaintiffs will have the ability to bring only meritorious cases to court, and defendants are more likely to resolve them faster.

**Main Research Questions**

In this work, we pose and examine three main research questions:
**RQ1**: *To what extent is it possible to determine the outcome of US class action cases using only the textual part of the complaints (without metadata)?*
**RQ2**: *To what extent can we use Temperature Scaling (TS) to better calibrate our models?*
**RQ3**: *To what extent can expert human lawyers solve the proposed task?*

**Contributions**

The contributions of this paper are four-fold:

- We curate a new specialized dataset of 10.8K class action complaints in the US from 2012 to 2022 annotated with the binary outcome: win or lose (plaintiff side). In contrast to most other LJP datasets it is (a) from a country with the common law system (where there are less datasets available), (b) it is specialized to class actions (important types of complaints ensuring justice for numerous often under-represented individuals), and (c) it uses the plaintiff's pleas as input instead of the facts, making the task more realistic. To the best of our knowledge, our work is the first dataset with plaintiff's pleas in the common law system and in the English language.
- We conduct a detailed analysis of the studied models using Integrated Gradients (IG) and model calibration using TS (Guo et al., 2017a).
- We perform an experiment with human experts on a randomly selected subset of the dataset, showing that our Longformer model both outperforms the human experts in terms of accuracy and calibration.
- We publicly release a sample of 3,000 cases from the annotated dataset[1] together with the human expert labels[2] and the code for the experiments[3].

## 2 Legal Background

### 2.1 Class Action Lawsuits

Class actions are a unique procedural instrument that allows one person to sue a company, not only on behalf of himself, but for everyone that has been injured by the same wrongdoing. In contrast to traditional lawsuits, in a class action lawsuit a plaintiff sues the defendant(s) on behalf of a class of absent parties. Class action lawsuits typically involve a minimum of 40 claimants. Rather than filing individual lawsuits for each damaged person, class actions allow the plaintiffs to unite and sue through a single proceeding. Thus, class actions are usually large and important cases and contain more complexity due to the high number of represented plaintiffs. These characteristics make class action a legal enforcement mechanism, along with police and regulators. Class actions both deter companies from harming people in the first place, and give compensation to the large number of victims hurt by the violation, giving consumers power over large corporations.

### 2.2 Definitions

**Civil Law vs. Common Law:** In both civil law and common law systems, courts rule based on laws and precedents (previous case law, mostly from the Supreme Court). However, in common law countries (mainly present in the UK and its former Colonies), case law dominates, whereas in civil law countries (most other countries) laws are more important. Note, that the differences are often not clear-cut, and courts usually use a combination of both laws and precedent for their rulings.
**Complaint:** A complaint is a written pleading to initiate a lawsuit. It includes the plaintiff's cause of action, the court's jurisdiction, and the plaintiff's demand for judicial relief. It is necessary for

[1] https://huggingface.co/datasets/darrow-ai/USClassActions
[2] https://huggingface.co/datasets/darrow-ai/USClassActionOutcomes_ExpertsAnnotations
[3] https://github.com/darrow-labs/ClassActionPrediction

the complaint to state all of the plaintiff's claims against the defendant, as well as what remedy the plaintiff seeks. A complaint must state "enough facts to state a claim to relief that is plausible on its face" (Twombly, 2007). The standards for filing a complaint vary from state to federal courts, or from one state to another. A typical class action complaint contains the allegations, the background details about both the plaintiff and the defendant, and the facts.

**Allegations:** In a complaint, allegations are statements of claimed facts. These statements are only considered allegations until they are proven. An allegation can be based on information and belief if the person making the statement is unsure of the facts. In the complaint, allegations can appear twice: once as a summary at the beginning and once in more detail later. There is usually a reference to the act that the plaintiff's attorney claims to have been violated in the allegations.

**Background Details:** The complaint contains background sections such as the plaintiff's history, class definitions, the defendant's history, and details about the platform/service in which the allegations took place.

**Plaintiff's Facts:** The plaintiff's facts or "factual background", are statements that can be proven and are often backed up with references and event dates. Note that the plaintiff's facts are written by the plaintiff lawyers.

**Facts Summary:** The facts summary or "factual description", are the summary of the accepted facts by the court and are written by the clerks or judges. The facts summary is usually more condensed in higher courts. Most previous LJP tasks used facts of this type. Since in this paper we consider complaints as input, when "facts" are mentioned we refer to the plaintiff's facts.

**Case Description:** The case description is written by the court clerks or judges and usually includes the header, the facts, the considerations, and the rulings.

### Class Action Outcomes

Table 1 shows the outcomes possible in class action cases. In the following, we briefly describe each of the outcomes.

**Settled:** "Settling a case" refers to resolving a dispute before the trial ends.

**Uncontested Dismissal:** Without any opposition from the parties, the case is dismissed and closed.

**Motion to Dismiss:** The case was dismissed by the court following the defendant's formal request for a court to dismiss the case.

| Outcome | Bin. Label | # Examples (%) |
|---|---|---|
| Settled | win | 5234 (48.64%) |
| Other - Plaintiff | win | 58 (00.52%) |
| Uncontested Dismissal | lose | 4544 (42.23%) |
| Motion to Dismiss | lose | 755 (07.01%) |
| Other - Defendant | lose | 170 (01.56%) |

Table 1: This table shows the original outcome together ruled by the court with the frequency and the final binarized label we map it to.

## 3 Related Work

LJP is an important and well-studied task in legal NLP. Cui et al. (2022) subdivide LJP into five subtasks: (a) In the *Article Recommendation Task*, systems predict relevant law articles for a given case (Aletras et al., 2016; Chalkidis et al., 2019; Ge et al., 2021). (b) The goal of the *Charge Prediction Task*, mainly studied in China, is to predict the counts the defendant is charged for based on the facts of the case (Zhong et al., 2018; Hu et al., 2018; Zhong et al., 2020). (c) In the *Prison Term Prediction Task*, systems predict the prison time for the defendant as ruled by the judge (Zhong et al., 2018; Chen et al., 2019). (d) In the *Court View Generation Task*, systems generate court views (explanation written by judges to interpret the judgment decision) (Ye et al., 2018; Wu et al., 2020). (e) In the *Plea Judgment Prediction Task*, systems predict the case outcome based on the case's facts (Niklaus et al., 2021b; Şulea et al., 2017; Lage-Freitas et al., 2022; Long et al., 2019; Ma et al., 2021; Strickson and De La Iglesia, 2020; Malik et al., 2021a; Alali et al., 2021). Since our work belongs to the PJP category, in the following, we elaborate more on the related work in this area.

**Civil Law** Niklaus et al. (2021b) released a trilingual (German, French, Italian) Swiss dataset from the Federal Supreme Court of Switzerland. They use the facts summary as input and predict a binary output: approval or dismissal of the plaintiff's pleas for approx. 85K decisions. Şulea et al. (2017) released a dataset of approx. 127K French Supreme Court cases. As input, they used the entire case description and not only the facts summary, presumably making the task considerably easier and a possible explanation for their high performance on the dataset. As output, they consider up to 8 classes of decisions ruled by the court. Lage-Freitas et al.

(2022) released a dataset comprising roughly 4K cases from a Brazilian State higher court (appellate court). They predicted three labels from the entire case description (written by the judges/clerks). Jacob de Menezes-Neto and Clementino (2022) release a large dataset of over 765K cases from the 5th Regional Federal Court of Brazil. They investigate a binary prediction task (whether the previous decision was reversed or not) using the entire case description as input. Long et al. (2019) studied the LJP task on 100K Chinese divorce proceedings considering three types of information as input: applicable law articles, fact description, and plaintiffs' pleas. Their model predicts a binary output. Ma et al. (2021) released a dataset comprising 70.5K civil cases (private lending) from China. They consider the more realistic task of inputting the plaintiff's complaints (together with debate data) instead of the easier facts summary used by most previous works. As output, their models predict three classes (reject, partially support and support). Similarly, our work also studies the more realistic (and challenging) use case of using the plaintiff's pleas as input instead of the heavily processed facts.

**Common Law** Strickson and De La Iglesia (2020) released a dataset of 5K cases from the UK's highest court of appeal. As input, they consider the case description and their models predict two labels (allow vs. dismiss). Malik et al. (2021a) study a dataset of 35K Indian Supreme Court cases in English. They use the case description as input and predict a binary outcome (accepted vs. rejected). Alali et al. (2021) study a dataset of 2.4K US Supreme Court decisions. Their models used the facts summary as input and predicted a binary output (first party won vs. second party won). In contrast, our dataset is $\sim$ 5 times larger and is specialized to the rare subset of class action cases.

Apart from Ma et al. (2021), the PJP task based on plaintiff's complaints has not been studied before. In contrast, most previous works studied textual input originating from the case description written by the court.

## 4 Dataset Description

In this section, we describe the dataset origin and statistics in detail. Additionally, we elaborate on the dataset construction process and the variants we produced.

Figures 2a and 2b show the distribution of cases across the most frequent states and courts in the dataset, respectively. Note that the origin of these class action lawsuits is very diverse, both across states and across courts. Not surprisingly, population-rich states like California, Florida, and New York lead the list. However, while California is more than double in population (39.5M vs. 20.2M as of April 2021), the number of class action lawsuits has the inverse relationship ($\sim$ 3K from New York and $\sim$ 1.8K from California). We assume that the complicated filing system in California could be a reason for this disparity[4].

### 4.1 Plaintiff's Pleas Instead of Facts Summary

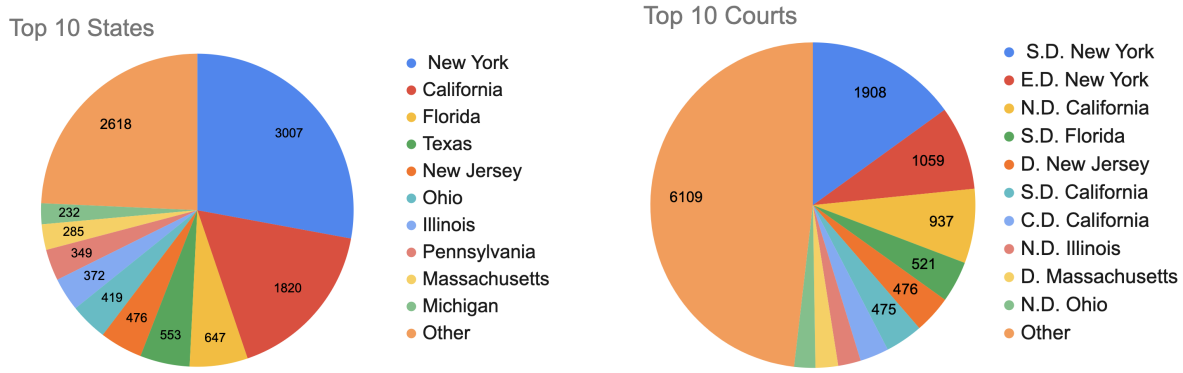
Condensing and extracting the relevant information from plaintiffs' pleas and court debates is a large part of the judge's work (Ma et al., 2021). This results in a condensed description of a case's facts. Most previous works consider this condensed description written by the judicial body (judges and clerks) as input. However, since a lot of qualified time has been spent on writing these descriptions, naturally, it makes the LJP task easier when using the court-written facts as input. Ma et al. (2021) were the first to consider the original plaintiff's pleas as input on Chinese data. In this work, to the best of our knowledge, we are the first to consider this harder task in the common law system (US class action cases in our case) and in the English language in general.

We do not consider the background details because our models might easily overfit on very specific data. In contrast, our goal was to create a dataset, where models need to focus on case-specific details to solve the task instead of being allowed to consider company-specific information such as number of employees or the area of business. We also disregard the introduction, containing metadata about the judge and the plaintiff.

### 4.2 Dataset Construction

To extract the plaintiffs' facts and allegations from each case, we manually reviewed hundreds of cases from different courts and different states to learn the structure of the document in each court to build a rule-based regex extraction system that detects the relevant text spans in each complaint. To summarize, constructing the dataset posed many technical difficulties due to the diverse nature of the

[4]Each court has its format of filing, and even courts within the same county do not usually use the same complaint filing format.




(a) Top 10 most frequent states

(b) Top 10 most frequent courts

Figure 2: Distribution of cases across states and courts.

complaint documents. At the preprocessing stage, we perform text cleaning, including removing some irrelevant text sections that our system incorrectly matched and removing duplicate sections.

### 4.3 Label Distribution

In this work we consider the task of binary legal judgment prediction. To do so, we simplified the labels. We used Table 1 to map the outcomes to either *win* or *lose* (for the plaintiff). After binarization the dataset is almost balanced with 5,469 (50.8%) *lose* cases and 5,290 (49.2%) *win* cases. Therefore, in our experiments, we just report the accuracy to keep it simple and make the scores more easily interpretable.

### 4.4 Dataset Variants

We experimented with different variants of the dataset to study the effect of the different parts of the text. We deliberately focused our attention more on the allegations because the facts contain a lot of repetitive content and are noisier than the allegations (many paragraphs only contain citations). Additionally, the facts contain many citations to laws, which are less relevant to the case's outcome according to domain experts (the facts are more generic and less case-specific than the allegations).

#### Full Text

The *Full Text* dataset combines the plaintiff's facts and the allegations but also disregards any background details. We concatenated the facts at the beginning and added the allegations parts to create one input text. We observe in Figure 3a that this dataset is rather long – almost 2700 tokens on average – with 10% of cases longer than 5400 tokens.

#### Unified Allegations

The *Unified Allegations* dataset consists of all case's allegations (mentioned in the complaint) concatenated together to form one input text . Approx. 2K documents did not contain any allegations (based on our extraction regexes), reducing the dataset size from 10.8K to 8.8K documents. The allegations make up a bit less than half of the full text complaint, as shown in Figure 3b (mean of ∼1,100 tokens and percentile 90 at ∼2,400 tokens).

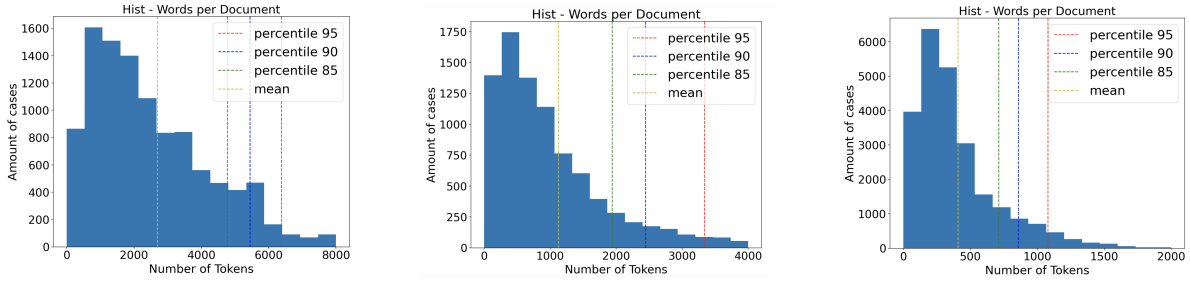
#### Separated Allegations

The *Separated Allegations* dataset considers each allegation as a separate sample, increasing the size from 8.8K to 25K documents. We considered this dataset to test whether the entire context is necessary. Figure 3c shows the length distribution for individual allegations. Surprisingly, even a single allegation can reach up to 2,000 tokens (∼ 4-5 pages of continuous text). However, most allegations (95%) are not longer than roughly 2 pages (1,100 tokens) with the average at 400 tokens.

## 5 Experiments

### 5.1 Experimental Setup

For all experiments, we truncated the text to the model's maximum sequence length (2,048 for Longformer and BigBird, 512 otherwise), unless otherwise specified. All experiments have been performed on the binarized labels (win or lose). We ran the experiments with 5-fold cross-validation and averaged across 5 random seeds. For more details regarding hyperparameter tuning and preprocessing, please refer to Appendix A.




(a) Full Text (b) Unified Allegations (c) Separated Allegations

Figure 3: Histograms for the three dataset variants (number of tokens calculated using bert-base-uncased tokenizer).

### 5.2 Methods

We compared the following pretrained transformer models: BERT (Devlin et al., 2019), LegalBERT (Chalkidis et al., 2020) (pretrained on diverse English legal data from Europe and the US with a domain-specific tokenizer), CaseLawBERT (Zheng et al., 2021) (pretrained on 37GB of US state and federal caselaw with a domain specific tokenizer), LegalRoBERTa[5] (continued pretraining from RoBERTa checkpoint on 4.6 GB of US caselaw and patents), BigBird (Zaheer et al., 2021) and Longformer (Beltagy et al., 2020). For all models, we used the publicly available base checkpoints on the Huggingface hub[6]. We ran our experiments with the Huggingface transformers library (Wolf et al., 2020) available under an Apache-2.0 license.

### 5.3 Results

Results are reported in the $mean_{\pm std}$ format averaged accuracy across 5 random seeds. Table 2 shows the main results. We observe that the setup considering the entire text is harder than when we only consider the allegations (best Full Text model is at $\sim$ 63% and worst allegations model is at $\sim$ 65%). These findings confirm our hypothesis, that the allegations encode more useful information than the facts (see Section 4.4) (the facts are often at the beginning of the complaints; thus the models on the Full Text dataset are likely to see mostly facts because of the truncation).

In line with previous findings (Chalkidis et al., 2021b, 2020; Zheng et al., 2021), models with legal pretraining outperform BERT also in our datasets (Unified Allegations and Separated Allegations). However, for LegalBERT the difference is small (only 0.5% above BERT). The models pretrained mostly or exclusively on US caselaw (LegalRoBERTa or CaseLawBERT respectively) perform better (up to 2% better than BERT), presumably because our dataset also originates from the US. CaseLawBERT achieves a much higher difference to BERT on the CaseHOLD task (4.6 F1) (Zheng et al., 2021) and on SCOTUS (7.6 macro-F1) (Chalkidis et al., 2021b). Both of these tasks are based on the same data as has been used in the pre-training of LegalRoBERTa and CaseLawBERT, whereas the complaints in our dataset are unseen by all models during pre-training. We suspect that this different data is the reason for the legal models not outperforming BERT as strongly as has been observed in other datasets.

| Method | Accuracy |
|---|---|
| Full Text (trunc. to 2048 tokens) | |
| Longformer | $62.87_{\pm 2.06}$ |
| BigBird | $63.26_{\pm 3.40}$ |
| Unified Allegations (trunc. to 512 tokens) | |
| BERT | $65.06_{\pm 1.67}$ |
| LegalBERT | $65.57_{\pm 0.26}$ |
| CaseLawBERT | $65.87_{\pm 0.60}$ |
| LegalRoBERTa | $65.95_{\pm 0.98}$ |
| Separated Allegations (trunc. to 512 tokens) | |
| BERT | $64.98_{\pm 1.08}$ |
| LegalBERT | $65.57_{\pm 0.62}$ |
| CaseLawBERT | $66.82_{\pm 0.78}$ |
| LegalRoBERTa | $65.97_{\pm 0.88}$ |

Table 2: Longformer and BigBird used a maximum sequence length of 2,048 tokens. All other models used 512 tokens. For all datasets, we truncated the text to fit the maximum sequence length.

## 6 Error Analysis

Neural Networks (NNs) and their latest incarnation, Transformers (Vaswani et al., 2017), work very well across a wide range of tasks, especially if

[5] https://huggingface.co/saibo/legal-roberta-base

[6] https://huggingface.co/models

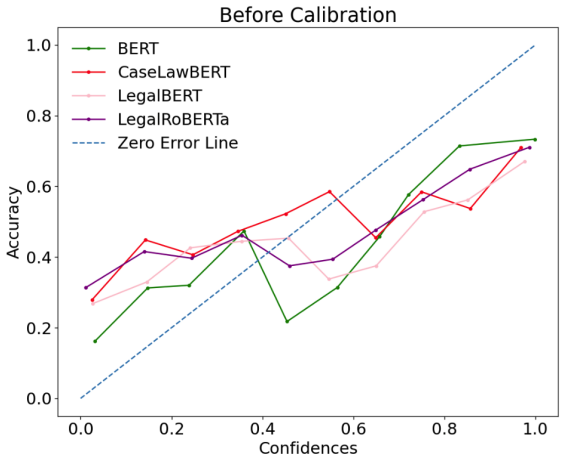
the tasks involve more "complicated" data like text or images. However, in contrast to traditional Machine Learning (ML) methods such as Linear Regression, they are not interpretable out-of-the-box. Neural Networks need additional methods to make them explain themselves better to humans. A rich body of literature investigates how to make NNs and especially Transformers more interpretable (Ribeiro et al., 2016; Sundararajan et al., 2017; Lundberg and Lee, 2017; Dhamdhere et al., 2018; Serrano and Smith, 2019; Bai et al., 2021). Interpretability is especially important in high-stakes domains such as law or medicine.

In the following two sections, we analyze our models using the two interpretability methods Calibration and IG to get a better understanding of their inner workings.

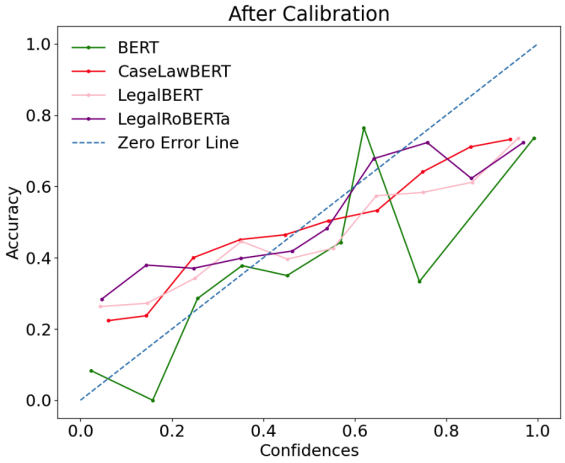
### 6.1 Calibration

In this section, we investigate to what extent our models are calibrated out-of-the-box and "calibratable". Calibration is a first step towards understanding whether the model output can be trusted (Guo et al., 2017b; Desai and Durrett, 2020): how aligned are the confidence scores with the actual empirical likelihoods? Thus, if the model assigns 60% probability to a label, then this label should be correct in 60% of cases if the model is calibrated. So, even if the model itself is a black-box, a calibrated model at least gives an indication whether it knows when it is wrong. This information can be very valuable when deploying models in the real world because it allows us to discard predictions where the model is below some certainty threshold. Well calibrated models are especially important in domains with high potential downside for users, such as predictive tools for court cases.

In this work, we follow Desai and Durrett (2020) by employing TS (Guo et al., 2017b) for calibrating our models using the netcal library[7] (Küppers et al., 2020) available under an Apache License 2.0 license. We show calibration plots in Figure 4 for BERT and the legal models on the Unified Allegations dataset and aggregated scores in Table 5 in Appendix B.3. We observe that the legal models are less calibrated than BERT before, but better calibrated after TS. So TS seems to calibrate domain-specific models better than general models. When comparing the calibration of our models with




(a) Before Calibration




(b) After Calibration

Figure 4: Calibration on the Unified Allegations dataset.

the calibration of models from the literature (Desai and Durrett, 2020), we note that our models are less calibrated overall (further away from the zero-error-line and higher ECE scores), both out-of-the-box and after applying TS. We hypothesize that the generally lower accuracy on our hard dataset also makes the models less calibrated, especially in the areas of high (> 0.8) and low (< 0.2) confidence. To the best of our knowledge, in legal NLP we are the first to perform such an analysis.

### 6.2 Integrated Gradients

We conduct a qualitative analysis of the LegalBERT model using IG[8] (Sundararajan et al., 2017) and show an illustrative example in Figure 5. We observe that the model focuses most on "flsa" an acronym for Fair Labor Standards Act[9] regulating

[7] https://github.com/fabiankueppers/calibration-framework

[8] https://github.com/cdpierse/transformers-interpret#sequence-classification-explainer

[9] https://www.dol.gov/agencies/whd/flsa

minimum wage and overtime among others. Further, the model focuses on "work" and "wages" possibly signaling a (limited) understanding of the connections between those concepts. Future work may investigate explainability of Pretrained Language Models (PLMs) in more detail on the LJP task.

## 7 Human Expert Annotations

Malik et al. (2021b) collected predictions for the judgment outcome of Indian Supreme Court cases from five legal experts. The experts agreed with the judges in 94% of the cases, on average. Note, however, that they have access to both the facts summary and the court's considerations. Their best model, XLNet + BiGRU, only achieves an accuracy of 78%. Contrarily, Jacob de Menezes-Neto and Clementino (2022) find that all their models outperform 22 highly skilled experts on LJP in Brazilian Federal Courts using the entire case description for prediction.

We asked legal experts (employees of our company) and US law students in their final year, to predict the judgment outcome of 200 randomly selected examples in our Full Text dataset. Note that they only had access to the facts and allegations from the plaintiff's pleas (same as our models), and not to the court case written by the judge. So, their task was much more difficult than the one posed to the annotators by Malik et al. (2021b) and Jacob de Menezes-Neto and Clementino (2022). In our task, participants (whether models or human experts) basically need to estimate how the court is going to decide based only on the plaintiff's pleas. For each document, our legal experts had to answer whether they think the plaintiff would win or lose the case. Furthermore, they also had to indicate their confidence level for being correct (from 1 – very unsure – to 5 – very sure). We made sure that the annotators did not look for any additional information regarding the complaint (e.g., news articles about the outcome or further information on different legal platforms) so that their answer is based only on the input text presented on the annotation platform. Figure 6 in Appendix C presents a screenshot of the annotation platform we used.

On the entire dataset sample (200 examples), the human experts achieve an accuracy of 53%. When we filtered out the samples where the human experts were not confident (confidence score 1, 2 or 3), they achieved an accuracy of 60%. The entire results for the human experts are shown in Appendix B.4 in Table 6. We also trained and evaluated a Longformer model for comparison with the human predictions. We randomly split our remaining dataset into 6,877 train and 1,851 validation examples. Surprisingly, the Longformer model outperforms the human expert predictions both on the entire annotated test dataset (63% vs. 53% Accuracy) and the dataset filtered for high human confidence (67% vs. 60% Accuracy). In contrast to the human experts, the Longformer model only had access to the first 2,048 tokens of the case. While the human performance increases more than the Longformer performance on the high-confidence dataset, the Longformer model also has a higher performance, suggesting that these cases are easier to predict.

The task proposed in our dataset seems very challenging, given that human experts face great challenges in solving it. Interestingly, on the Indian dataset the humans clearly outperform the models, whereas in the Brazilian dataset it is reversed, similar to our results. Note that lawyers are often specialized in very narrow domains (legal areas). The cases in our dataset may be very diverse, and thus a generic model might be better suited for this task than specialized human experts. Future work may investigate this finding in more detail.

Figure 1 shows the calibration plot on the Full Text dataset, comparing Longformer before and after calibration with the human confidence scores. We observe that Longformer is already well calibrated in comparison to the human experts. Using TS, the Expected Calibration Error (ECE) of Longformer can be reduced from 5.14 to 2.34, whereas the ECE of the human experts lies at 17.5. Again, as mentioned in Section 6.1, the lower accuracy of the humans might explain their worse calibration compared to Longformer.

## 8 Conclusions and Future Work

### Answers to the Research Questions

**RQ1**: *To what extent is it possible to determine the winner of US class action cases using only the textual part of the complaints (without metadata)?* It is possible, to some extent, to determine the winner of US class action cases using only the textual part of the complaints. Our best model achieves an accuracy of 66.8% (LegalRoBERTa) on the datasets using only the allegations. However, as this number shows, there is still a lot of room for improvement.

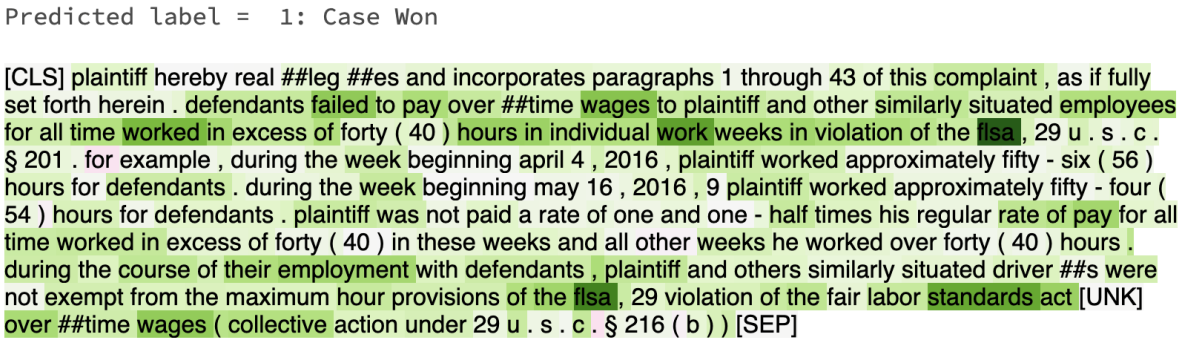



Figure 5: Analysis using Integrated Gradients (IG)

**RQ2**: *To what extent can we use Temperature Scaling (TS) to better calibrate our models?* Similar to Natural Language Inference, Paraphrase Detection and Commonsense Reasoning tasks (Desai and Durrett, 2020), we also find that in the PJP task, TS helps in calibrating pretrained transformers. In our best model, TS led to a decrease in ECE scores from 28 to 2.

**RQ3**: *To what extent can expert human lawyers solve the proposed task?* Expert human lawyers perform better than chance on a randomly selected dataset of 200 samples and can increase their accuracy from 53% to 60% when they are confident in their decision. However, they are still outperformed by a Longformer model having access to only the first 2,048 tokens in both scenarios.

## Conclusions

We release a challenging new dataset of class action lawsuits for the more realistic PJP task (where the input is based on the complaints instead of the further processed facts summary written by the judge) in the US, a jurisdiction with the common law system. Additionally, we calibrated our models using TS and found that despite the relatively low accuracy (66% for the best model), relatively low ECE scores around 2 can be achieved. Finally, we find that our Longformer model is 10% more accurate than the human experts on our dataset despite having only access to the first 2,048 tokens of the case.

## Limitations

Our best model achieves an accuracy of 66%. This may suggest that either the task posed in this dataset is very hard, or we did not optimize our models enough. The results achieved by the human experts suggests that the former is the case. However, we believe much more work is needed here.

Although we did some first efforts to interpret our model's outputs using Calibration and IG, the literature knows a host of other explainability methods (Molnar, 2022). We leave a more thorough qualitative analysis involving domain experts and explainability methods for future work.

Our experiments were performed only on relatively short input spans (512 tokens for allegations, and 2048 for full text). Longformer or BigBird support input spans until 4096 tokens. Another possibility is the use of hierarchical models, as employed for example by Niklaus et al. (2022); Dai et al. (2022) that can also easily scale to 4096 tokens given the right hardware. With 4096 tokens, we could fully encode all allegations and almost 80% percent of the full texts. We leave these investigations to future work.

## Future Work

Since the legal models outperformed BERT only to a small margin, we suspect that further pretraining (Gururangan et al., 2020) on in-domain data might further enhance the performance. Additionally, in future work, we plan to study the domain-specific PJP and whether domain-specific models are better than generic model or human experts.

Large PLMs have proved to be very strong few shot learners in many tasks (Brown et al., 2020; Chowdhery et al., 2022). The use of such models may bring performance boosts also in our studied task. We leave experimentation using different prompting strategies for future work (Arora et al., 2022; Wei et al., 2022; Suzgun et al., 2022).

We discovered through our analysis using IG that some legal domains have a strong correlation to a particular label. To produce complaints with a higher success likelihood in court, future studies may examine the linguistic structure of successful allegations.

## Ethics Statement

The goal of this research is to achieve a better understanding of LJP to broaden the discussion and aid practitioners in developing better technology for both legal experts and non-specialists. We believe that this is a crucial application area, where research should be done (Tsarapatsanis and Aletras, 2021) to improve legal services and democratize legal data, making it more accessible to end-users, while also highlighting (informing the audience on) the various multi-aspect deficiencies seeking a responsible and ethical (fair) deployment of legal-oriented technology.

In this direction, we study how we can best build our dataset to maximize accuracy of our models on the task. Additionally, we study the inner workings of the models using Integrated Gradients and make sure that our models are calibrated. A well calibrated model outputs confidence probabilities in line with actual likelihoods, thus giving the users the possibility of discarding low-confidence predictions or at least treating them with caution.

Lawyers often perform the LJP task by giving their clients advice on how high the chances for success are in court for specific cases. Given the complaint documents, we were able to show in this work that our models outperformed human experts in this task.

But, like with any other application (like content moderation) or domain (e.g., medical), reckless usage (deployment) of such technology poses a real risk. According to our opinion, comparable technology should only be used to support human specialists (legal scholars, or legal professionals).

## Acknowledgements

We thank all the anonymous reviewers for their insightful comments. We thank the two employees at Darrow for the annotation of the dataset.

## References

Mohammad Alali, Shaayan Syed, Mohammed Alsayed, Smit Patel, and Hemanth Bodala. 2021. JUSTICE: A Benchmark Dataset for Supreme Court's Judgment Prediction. ArXiv:2112.03414 [cs].

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93. Publisher: PeerJ Inc.

I. Angelidis, Ilias Chalkidis, and M. Koubarakis. 2018. Named Entity Recognition, Linking and Generation for Greek Legislation. In *JURIX*.

Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask Me Anything: A simple strategy for prompting language models. ArXiv:2210.02441 [cs].

Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. 2021. Why Attentions May Not Be Interpretable? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 25–34, New York, NY, USA. Association for Computing Machinery.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*. ArXiv: 2004.05150.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. ArXiv: 2005.14165.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021a. MultiEURLEX – A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv:2109.00904 [cs]*. ArXiv: 2109.00904.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. *arXiv:2010.02559 [cs]*. ArXiv: 2010.02559.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael James Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021b. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. SSRN Scholarly Paper ID 3936759, Social Science Research Network, Rochester, NY.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-Based Prison Term Prediction with Deep Gating Network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

*Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6362–6367, Hong Kong, China. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv:2204.02311 [cs]*. ArXiv: 2204.02311.

Junyun Cui, Xiaoyu Shen, Feiping Nie, Z. Wang, Jinglong Wang, and Yulong Chen. 2022. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *ArXiv*, abs/2204.04859.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting Transformer-based Models for Long Document Classification. *arXiv:2204.06683 [cs]*. ArXiv: 2204.06683.

Shrey Desai and Greg Durrett. 2020. Calibration of Pre-trained Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. 2018. How Important Is a Neuron? ArXiv:1805.12233 [cs, stat].

Jidong Ge, Yunyun huang, Xiaoyu Shen, Chuanyi Li, and Wei Hu. 2021. Learning Fine-grained Fact-Article Correspondence in Legal Cases. ArXiv:2104.10726 [cs].

Ingo Glaser and Florian Matthes. 2020. Classification of German Court Rulings: Detecting the Area of Law. page 10.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017a. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017b. On Calibration of Modern Neural Networks. Number: arXiv:1706.04599 arXiv:1706.04599 [cs].

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv:2004.10964 [cs]*. ArXiv: 2004.10964.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. ArXiv:2103.06268 [cs].

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-Shot Charge Prediction with Discriminative Legal Attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Elias Jacob de Menezes-Neto and Marco Bruno Miranda Clementino. 2022. Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts. *PLOS ONE*, 17(7):e0272287.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A Corpus for Automatic Summarization of US Legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Fabian Küppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. 2020. Multivariate confidence calibration for object detection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

André Lage-Freitas, Héctor Allende-Cid, Orivaldo Santana, and Lívia Oliveira-Lage. 2022. Predicting Brazilian Court Decisions. *PeerJ Computer Science*, 8:e904. Publisher: PeerJ Inc.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-Grained Named Entity Recognition in Legal Documents. In Maribel Acosta, Philippe Cudré-Mauroux, Maria Maleshkova, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter, editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, volume 11702, pages 272–287. Springer International Publishing, Cham.

Shangbang Long, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Automatic Judgment Prediction via Legal Reading Comprehension. In *Chinese Computational Linguistics*, Lecture Notes in Computer Science, pages 558–572, Cham. Springer International Publishing.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pedro Henrique Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In *Computational Processing of the Portuguese Language*, Lecture Notes in Computer Science, pages 313–323, Cham. Springer International Publishing.

Luyao Ma, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Wei Ye, Changlong Sun, and Shikun Zhang. 2021. Legal Judgment Prediction with Multi-Stage CaseRepresentation Learning in the Real Court Setting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 993–1002. ArXiv:2107.05192 [cs].

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021a. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021b. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Christoph Molnar. 2022. *Interpretable Machine Learning*, 2 edition.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021a. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021b. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. ArXiv:2209.12325 [cs].

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-LexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities. ArXiv:2206.10883 [cs].

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal Judgement Prediction for UK Courts. In *Proceedings of the 2020 The 3rd International Conference on Information Science and System*, ICISS 2020, pages 204–209, New York, NY, USA. Association for Computing Machinery.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. ArXiv:1703.01365 [cs].

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. ArXiv:2210.09261 [cs].

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

550 U.S. at 570 Twombly. 2007. Bell atlantic corp. v. twombly. *Justia*.

Stefanie Urchs, Jelena Mitrović, and Michael Granitzer. 2022. Design and Implementation of German Legal Decision Corpora. pages 515–521.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*. ArXiv: 1706.03762.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903 [cs].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Marco Wrzalik and Dirk Krechel. 2021. GerDaLIR: A German Dataset for Legal Information Retrieval. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 123–128, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-Biased Court's View Generation with Causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780, Online. Association for Computational Linguistics.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big Bird: Transformers for Longer Sequences. *arXiv:2007.14062 [cs, stat]*. ArXiv: 2007.14062.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. *arXiv:2104.08671 [cs]*. ArXiv: 2104.08671 version: 3.

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. Association for Computational Linguistics.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively Questioning and Answering for Interpretable Legal Judgment Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1250–1257. Number: 01.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the Law Area and Decisions of French Supreme Court Cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria. INCOMA Ltd.

# A Additional Training Details

## A.1 Hyperparameter Tuning

We randomly split the data into 70% train, 15% validation and 15% test split. We searched the learning rate in {1e-6, 5e-5, 1e-5} and had the best results with 1e-5. We searched dropout in {0, 0.001, 0.1, 0.2} and finally chose 0. We searched the batch size in {16, 32, 64} and chose 16. Where GPU memory was not sufficient, we used gradient accumulation for a total batch size of 16. We searched the activation function in {Relu, SoftMax, LeakyRelu} and chose SoftMax. We searched weight decay in {0, 0.1} and found 0 to perform best. We used AMP mixed precision training and evaluation to reduce costs. We used early stopping on the validation loss with patience 2. If early stopping was not invoked, we trained for a maximum of 10 epochs. We used an AWS EC2 G5 instance with 4 CPU cores, 16 GB RAM and one NVIDIA A10G GPU (24 GB of GPU memory)

## A.2 Preprocessing

We experimented with the following preprocessing methods: (a) removing punctuation; (b) removing numerals; (c) stemming; (d) lemmatization; and (e) entity masking (e.g., "Plaintiff James won would receive 30% from the 3 million compensation fund" → "PERSON won would receive PERCENT from the MONEY compensation fund"). We found that only stemming improved the results.

| Method | Max Seq Len | Accuracy |
|---|---|---|
| Full Text | | |
| Longformer | 2048 | $63.64_{\pm 0.72}$ |
| BigBird | 2048 | $62.00_{\pm 1.08}$ |
| Separated Allegations | | |
| BERT | 512 | $64.82_{\pm 1.73}$ |
| CaseLawBERT | 512 | $66.06_{\pm 0.84}$ |
| LegalBERT | 512 | $64.57_{\pm 1.89}$ |
| LegalRoBERTa | 512 | $65.41_{\pm 1.09}$ |

Table 3: Longformer and BigBird used a maximum sequence length of 2,048 tokens. All other models used 512 tokens. For all datasets, we filtered out the rows larger than the maximum sequence length.

### A.3 Training Times

On the Unified Allegations dataset, training took approximately one hour for all the investigated models. On the Separated Allegations dataset, it took approximately two hours per model. On the Full Text dataset, it took approximately six hours for Longformer and approximately eight hours for BigBird. All training times are counted for five folds and one random seed on an AWS EC2 G5 instance with 4 CPU cores, 16 GB RAM and one NVIDIA A10G GPU (24GB of GPU memory).

### A.4 Library Versions

We used the following libraries and associated versions: python 3.8, transformers 4.17.0, xgboost 1.5.2, torch 1.11.0+cu113, tokenizers 0.12.1, spacy 3.2.3, scikit-learn 1.1.1, pandas 1.3.4, numpy 1.20.3, netcal 1.2.1, nltk 3.6.5, optuna 2.10.1, matplotlib 3.4.3.

## B Additional Results

### B.1 Filtering the Datasets

In Table 3 we show results for the Filter setup, where we filtered out texts containing more tokens than the maximum sequence lengths of the models used. We note that the results don't change significantly in comparison to Table 2 (Truncation setup).

### B.2 XGBoost

Table 4 shows the results for using XGBoost (Chen and Guestrin, 2016) on top of the embeddings instead of simple linear layers as it is reported in Table 2. We observe that this more sophisticated classification layer does not improve results.

| Method | Max Seq Len | Accuracy |
|---|---|---|
| Full Text | | |
| BERT | 512 | $60.40_{\pm 0.90}$ |
| LegalBERT | 512 | $61.79_{\pm 1.13}$ |
| CaseLawBERT | 512 | $60.65_{\pm 0.32}$ |
| LegalRoBERTa | 512 | $60.37_{\pm 0.66}$ |
| Longformer | 2048 | $59.96_{\pm 1.24}$ |
| BigBird | 2048 | $60.98_{\pm 0.70}$ |
| Unified Allegations | | |
| BERT | 512 | $62.08_{\pm 0.71}$ |
| LegalBERT | 512 | $63.01_{\pm 0.60}$ |
| CaseLawBERT | 512 | $62.22_{\pm 0.59}$ |
| LegalRoBERTa | 512 | $62.32_{\pm 1.12}$ |
| Longformer | 512 | $61.7_{\pm 0.82}$ |
| BigBird | 512 | $61.13_{\pm 1.02}$ |
| Separated Allegations | | |
| BERT | 512 | $63.19_{\pm 0.49}$ |
| LegalBERT | 512 | $64.17_{\pm 0.44}$ |
| CaseLawBERT | 512 | $63.81_{\pm 0.67}$ |
| LegalRoBERTa | 512 | $64.52_{\pm 0.30}$ |
| Longformer | 512 | $64.65_{\pm 0.40}$ |
| BigBird | 512 | $63.38_{\pm 0.31}$ |

Table 4: We fed the embeddings of the transformer models into an XGBoost (Chen and Guestrin, 2016). For all datasets, we truncated the text to fit the maximum sequence length.

### B.3 Calibration Results

Table 5 shows the detailed aggregated ECE scores together with the optimal temperature and the accuracy on the Unified Allegations dataset.

### B.4 Human Results

Table 6 shows the results of the human experts on the 200 randomly selected examples.

## C Annotation Platform

Figure 6 shows a screenshot of the annotation platform our human experts used.

## D Example Complaint

Figures 7 and 8 show an example of a complaint present in the dataset.

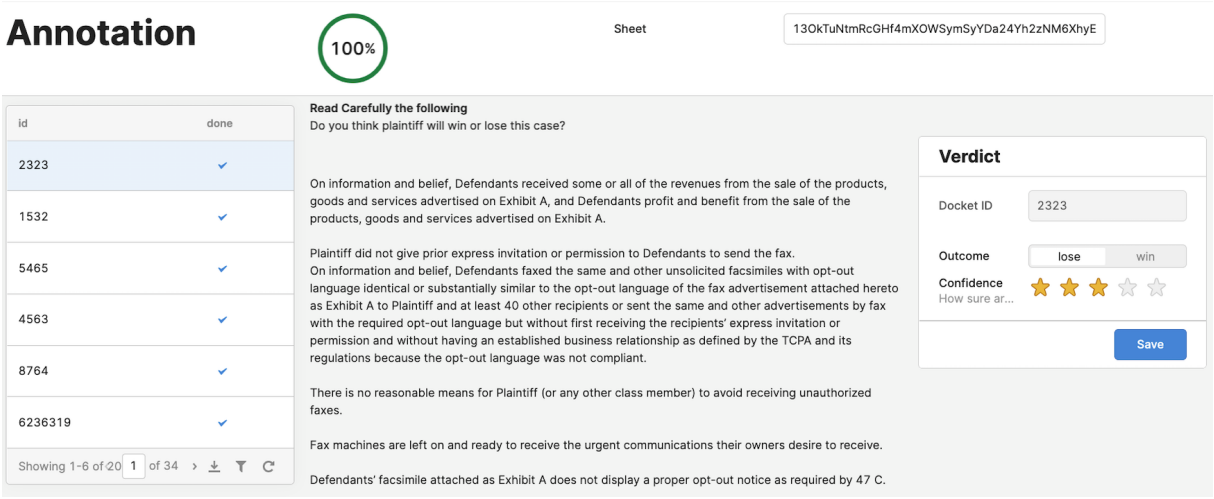



Figure 6: The platform for the human annotations.

| Method | Opt. Temp. | ECE Before | ECE After | Accuracy |
|---|---|---|---|---|
| BERT | $0.19_{\pm 0.03}$ | $23.44_{\pm 3.20}$ | $5.06_{\pm 1.96}$ | $65.06_{\pm 1.67}$ |
| CaseLawBERT | $0.20_{\pm 0.03}$ | $25.67_{\pm 2.32}$ | $2.59_{\pm 0.90}$ | $65.57_{\pm 0.60}$ |
| LegalBERT | $0.22_{\pm 0.02}$ | $24.78_{\pm 1.13}$ | $3.06_{\pm 1.78}$ | $65.87_{\pm 0.26}$ |
| LegalRobertaBase | $0.13_{\pm 0.02}$ | $28.02_{\pm 2.16}$ | $1.92_{\pm 0.85}$ | $65.95_{\pm 0.98}$ |

Table 5: Calibration results on the Unified Allegations dataset. The text was always truncated to fit the model's maximum sequence length of 512 tokens. Opt. Temp. abbreviates the optimal temperature used for calibrating the models.

| | Precision | Recall | F1-score | # Examples |
|---|---|---|---|---|
| All Results | | | | |
| lose | 49.41 | 45.65 | 47.45 | 92 |
| win | 56.52 | 60.18 | 58.29 | 108 |
| accuracy | - | - | 53.50 | 200 |
| High Confidence | | | | |
| lose | 75.00 | 37.50 | 50.00 | 24 |
| win | 54.54 | 85.71 | 66.66 | 21 |
| accuracy | - | - | 60.00 | 45 |

Table 6: Results of the human experts on the 200 randomly selected cases. Under High Confidence we show the results for only the examples where the human experts rated their confidence at 4 or 5 out of 5.

**IN THE UNITED STATES DISTRICT COURT**
**FOR THE NORTHERN DISTRICT OF ILLINOIS**
**EASTERN DIVISION**

| | | |
|---|---|---|
| ANTHONY HALL, on behalf of himself and all others similarly situated, | ) | |
| Plaintiff, | ) | Case No. 20-cv-00846 |
| vs. | ) | |
| CLEARVIEW AI, INC., and. CDW GOVERNMENT LLC; | ) | Jury Demanded |
| Defendants. | ) | |

**CLASS ACTION COMPLAINT**

Plaintiff Anthony Hall, on behalf of himself and a putative class ("Plaintiff" or "Hall"), brings this Class Action Complaint against Defendants Clearview AI, Inc ("Clearview"); CDW Government, LLC ("CDW") and alleges the following:

**Introduction**

1. A New York Times article published on January 18, 2020 introduced Americans to the then relatively unknown company Clearview AI, Inc. The article described a dystopian surveillance database, owned and operated by a private company and leased to the highest bidder.

2. Clearview AI's database includes the photographs, and personal and private data, including names, home addresses, and work addresses, of millions of Americans. Clearview acquired the billions of data points by "scraping" or harvesting the data from publicly available internet-based platforms such as Facebook, Instagram, and Twitter.



3. But Clearview's database is unique – it has run every one of the 3 billion photographs it has acquired through facial recognition software to extract and index the unique biometric data from each face. The database thus also contains the biometric identifiers and information of millions of Americans. Any private citizen can be identified by uploading a photo to the database. Once identified, the end-user then has access to all of the individual's personal details that Clearview has also obtained.

4. A second article published in the Chicago Sun-Times on January 29, 2020 revealed that the Chicago Police Department was using Clearview's surveillance database to aid in law enforcement operations.

**Jurisdiction**

5. This Court has jurisdiction under 28 U.S.C. § 1332(d)(2), the Class Action Fairness Act ("CAFA") because there are 100 or more members of the class, the parties and putative class members are minimally diverse and the aggregate amount in controversy is greater than $5,000,000.

6. This Court has personal jurisdiction over Clearview because they conduct a substantial amount of business here which forms the basis of Plaintiffs' claims. Clearview has made their surveillance database, which contains the private and personal data and biometric information of thousands of Illinois residents, available to Chicago Police department. All defendants' violations of Illinois law are based on and arise from their contacts with the state and its residents. The court has personal jurisdiction over CDW because they are an Illinois company headquartered in Illinois.

7. Venue is proper here under 28 U.S.C. § 1391(b)(2) because a substantial amount of the acts and omissions giving rise to the claims occurred in Illinois.

Figure 7: These are the first two pages from an example complaint.



80. Plaintiff and the Class seek:

a. $1,000 for the Plaintiff and each member of the class for each and every separate negligent violation;

b. $5,000 for the Plaintiff and each member of the class for each and every separate intentional or reckless violation;

c. punitive damages;

d. costs, expenses, and reasonable attorneys' fees;

e. and, any other relief this court deems proper.

**COUNT III – ILLINOIS CONSUMER FRAUD AND UNFAIR BUSINESS PRACTICES ACT – CLEARVIEW AND CDW**

81. At all times relevant, Defendants were engaged in trade or commerce in the state: Clearview and CDW leased, sold, or otherwise provided, for profit, access to the surveillance database to agencies within Illinois such as the CPD.

82. At all times relevant, Plaintiff and members of the class were consumers within the meaning of ICFA.

83. Defendants practice of unauthorized scraping or harvesting of Plaintiff's and the Class members' photos, videos, private and personal information, and its conversion into biometric information and identifiers to add to their surveillance database is an unfair practice.

84. This practice has caused substantial injury and harm to Plaintiff and the members of the Class. It has also forced the Plaintiff to retain counsel to force Clearview to comply with BIPA and redress other violations of state law.

85. Plaintiff and the Class seek:

a. actual damages;



b. punitive damages;

c. costs, expenses, and reasonable attorneys' fees;

d. and, any other relief this court deems proper.

**COUNT IV – CONVERSION – CLEARVIEW AND CDW**

86. Plaintiff and each Class member have a personal property right in their biometric information and identifiers.

87. Defendants assumed control over the biometric information and identifiers of Plaintiff and the Class with their knowledge or authorization. Defendants' actions impaired Plaintiff and Class members' exclusive right to control their property.

88. Plaintiff and the Class seek:

a. the greater of actual damages or the profits gained by CDW and Clearview from the conversion of Plaintiff and Class members property;

b. punitive damages;

c. and, any other relief this court deems proper.

**Jury Demand**

Plaintiff demands a trial by jury.

February 5, 2020

[Signature Page Follows]

Figure 8: These are the last two pages from an example complaint.