Deconfounding Legal Judgment Prediction for European Court of Human Rights Cases Towards Better Alignment with Experts

Santosh T.Y.S.S[1*] and **Shanshan Xu**[1*] and **Oana Ichim**[2] and **Matthias Grabmair**[1]
[1]School of Computation, Information, and Technology; Technical University of Munich, Germany
[2]Graduate Institute of International and Development Studies, Geneva, Switzerland
{santosh.tokala, shanshan.xu, matthias.grabmair}@tum.de
oana.ichim@graduateinstitute.ch

Abstract

This work demonstrates that Legal Judgement Prediction systems without expert-informed adjustments can be vulnerable to shallow, distracting surface signals that arise from corpus construction, case distribution, and confounding factors. To mitigate this, we use domain expertise to strategically identify statistically predictive but legally irrelevant information. We adopt adversarial training to prevent the system from relying on it. We evaluate our deconfounded models by employing interpretability techniques and comparing to expert annotations. Quantitative experiments and qualitative analysis show that our deconfounded model consistently aligns better with expert rationales than baselines trained for prediction only. We further contribute a set of reference expert annotations to the validation and testing partitions of an existing benchmark dataset of European Court of Human Rights cases.

1 Introduction

The task of Legal Judgment Prediction (LJP) has recently gained increasing attention in the legal and mainstream NLP communities (Aletras et al., 2016; Zhong et al., 2018; Medvedeva et al., 2020; Liu et al., 2019; Sert et al., 2021). Legal cases are resolved through the exchange of arguments in front of a decision body by lawyers who represent litigating parties. This typically involves evidential reasoning, the determination of relevant rules from sources of law (e.g., codes, regulations, precedent), their application to the case, and the balancing of legal and societal values. In the NLP context, LJP takes the form of classifying the outcome of a case from some textual representation of its specific facts, effectively skipping legal reasoning. This forms a counterpoint to knowledge-focused approaches to outcome prediction (e.g., Brüninghaus and Ashley, 2005; Branting, 2013; Grabmair, 2017) that connect to a lawyer's understanding of the domain but also require substantial knowledge engineering.

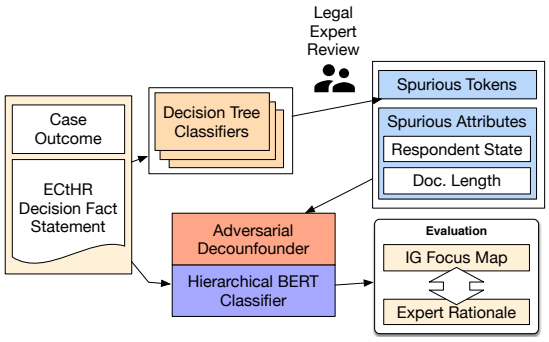


Figure 1: Our deconfounding experiment architecture

This carries particular risk in the legal domain, where systems may rely on data elements that are statistically predictive but legally irrelevant, or even forbidden as decision criteria (e.g., the race of an accused person). This can lead to undesirable consequences, ranging from suboptimal litigation strategy decisions, flawed inference about factors predictive for the outcome, to disparate impact of decisions across groups that are to be treated equally. If legal decisions are to be informed by predictive systems processing textual case descriptions, then such systems must strive to be as closely aligned with legally relevant and permissible parts of the input as possible.

In this work, we focus on LJP for the European Court of Human Rights (ECtHR), which adjudicates complaints by individuals against states about alleged violations of their rights as enshrined in the European Convention of Human Rights. We trained deep neural models on four tasks across two existing, related datasets (Chalkidis et al., 2019, 2022a) around predicting such violations alleged by the claimant and decided by the court. We find that the models substantially base their predictions on aspects of the text that correlate with the outcome

[*]These authors contributed equally to this work

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1120–1138
December 7-11, 2022

but either have no legal bearing or are forbidden nationality-related information that stem from the distribution of cases arising at the court.

To improve the alignment of model focus with legal expert understanding, we apply a series of deconfounding measures, including a vocabulary-based method which identifies predictive tokens using a simple model. The third author, who is an ECtHR expert, then identifies distractors among them. The distracting signal can subsequently be removed from the encodings via adversarial training. This procedure is an effective way of engaging with domain experts and obtaining information about what the model should be steered away from by means of deconfounding, rather than trying to attract the model towards relevant elements via expensive data collection for supervised training. For simplicity, throughout this paper, we use 'deconfounding' in an inclusive sense as the mitigation of distracting effects of (a) confounders in the statistical sense that influence both the dependent and independent variables, (b) reverse causation relationships, and (c) other attributes that spuriously correlate with the target variable. See Fig. 1 for an overview of our experiment design.

We evaluate our trained and deconfounded models with regard to an alignment of its explanation rationales with (1) a dataset of expert passage relevance assessments we collected and will make available to community as a supplement to Chalkidis et al. (2019), and (2) on expert relevance assessments published as part of Chalkidis et al. (2021). Our results show that our deconfounding steps succeed in improving the model focus alignment with expert-identified, relevant patterns on both sets of reference annotations.

In sum, we make the following contributions:

- We introduce an expert-informed deconfounding method which identifies distracting effects from confounders and spurious correlations using a simple model, and mitigates them through adversarial training, thus helping to improve the alignment of the model focus with legal expert rationales.
- We empirically evaluate this method on four tasks in legal judgment prediction on ECtHR data and show that our model consistently aligns better with expert rationales than a baseline trained for the prediction target only.
- We release a set of gold rationales annotated by an ECtHR expert as a supplement to an existing dataset to facilitate future work on deriving more useful insight from trained predictive systems in the legal domain.[*]

2 Related Work

LJP as an NLP task has been tackled using n-gram representations (e.g., Aletras et al., 2016; Medvedeva et al., 2020), word embeddings and domain models (Branting et al., 2021), and deep neural networks (e.g., Chalkidis et al., 2019; Ma et al., 2021; Xu et al., 2020). Special attention must be given to the origin of the text from which the prediction is to be made. Medvedeva et al. (2021, 2022) recharacterize LJP on texts produced before the outcome is known as 'forecasting' and observes that most current works 'classify' judgments based on the data compiled after the outcome has been determined. They also find that forecasting is a harder task. This result is consistent with our finding of confounding effects from text production by the ECtHR, resulting in a prediction from fact descriptions that were influenced by the decision.

Moverover, the relationship between the information LJP models rely on and legal expert analysis of texts remains underexplored. Bhambhoria et al. (2021) find that transformer-based models exploit spurious correlations and that simple models, such as XGBoost, can achieve similar performance. Chalkidis et al. (2021) extract model rationales for alleged violation prediction and observes limited overlap with expert markup. Similarly, a small study in Branting et al. (2021) finds that users do not perceive case prediction-derived highlighting as useful in making predictions themselves. Our work contributes to this state of the art by using adversarial deconfounding to improve the overlap between what systems predict from with what legal experts consider relevant.

Deconfounding A growing number of works have raised awareness that deep neural models may exploit spurious statistical patterns and take erroneous shortcuts (McCoy et al., 2019; Bender and Koller, 2020; Geirhos et al., 2020). A common method of mitigating this is adversarial learning. Pryzant et al. 2018 use a gradient reversal layer (Ganin et al., 2016) to deconfound lexicons in text classification. Other domains that adopt adversarial training to

[*]Our rationales and code are available at https://github.com/TUMLegalTech/deconfounding_echr_emnlp22

eliminate confounders include bioinformatics (Dincer et al., 2020) and political science (Roberts et al., 2020). Many existing works on identifying shortcuts focus on situations where these patterns are known in advance and may require potentially expensive data collection. In fairness-focused legal NLP, Chalkidis et al. (2022b) observe and remedy group disparities in LJP performance on the ECtHR informed by metadata attributes (respondent state, applicant gender, applicant age). We extend this to explainability in LJP by involving a legal expert in a procedure that allows an efficient, incremental identification of distracting information, as well as its removal via adversarial training.

Interpretability We employ interpretability techniques to evaluate model alignment with expert rationales. Danilevsky et al. (2020) reviews and categorizes the main current interpretability methods. Though initial works (Ghaeini et al., 2018; Lee et al., 2017) used attention scores as explanation for model decisions, Bastings and Filippova (2020); Serrano and Smith (2019) point out that saliency methods, such as gradient based methods (Sundararajan et al., 2017; Li et al., 2016), propagation based methods (Bach et al., 2015), occlusion based methods (Zeiler and Fergus, 2014), and surrogate model based methods (Ribeiro et al., 2016) are better suited for explainability analysis. However, the reliability and informativeness of these methods remains an open research problem. Our model uses the currently most commonly used Integrated Gradients (IG) (Sundararajan et al., 2017), which computes the gradient of the model's output with respect to its input features.

3 ECtHR Tasks & Datasets

The ECtHR has been the subject of substantial prior work in LJP. We use two datasets for model training and evaluation: First, for **binary violation** we use the dataset by Chalkidis et al. (2019) of approx. 11k case fact statements, where the target is to predict whether the court has found at least one convention article to be violated. To evaluate alignment, we annotate 50 (25 each) expert rationales for cases from both the development and test partitions (See App. C for the annotation process). Second, for **article-specific violation**, we use the LexGLUE dataset by Chalkidis et al. (2022a), which consists of 11k case fact statements along with information about which convention articles have been alleged to be violated, and which the court has found to be violated. For alignment, we merge this data with the 50 test set rationales from Chalkidis et al. (2021). While both datasets stem from the ECtHR's public database, they differ in case facts and outcome distribution as we explain in Sec. 3.1. The input texts consist of each case's FACTS section extracted from ECtHR judgments. This section is drafted by court staff over the course of the case proceedings. While it does not contain the outcome explicitly, it is not finalized before the final decision has been determined, potentially creating confounding effects.

We conduct experiments on four LJP tasks:

Task J - Binary Violation For our task **J**, the model is given a fact statement and is asked to predict whether or not any article of the convention has been violated. We train our models on Chalkidis et al. (2019) and evaluate alignment on the set of expert rationales we collected.

Task B - Article Allegation We train and evaluate on LexGLUE's *ECtHR B*,[*] where the fact description is the basis to predict the set of convention articles that the claimant alleges to have been violated. It can be conceptualized as topic classification in that the system needs to identify suitable candidate articles (e.g., the right to respect for private and family life) from fact statements (e.g., about government surveillance). We test alignment on the expert rationales by Chalkidis et al. (2021).

Task A - Article Violation We also experiment with LexGLUE's *ECtHR A*, which is to predict which of the convention's articles has been deemed violated by the court from a case's fact description. Task A is a more difficult version of task B, where both an identification of suitable articles and a prediction of their violation must be performed. For alignment, we again use the expert rationales by Chalkidis et al. (2021), which are technically intended for task *ECtHR B*, but which we consider to also be suitable for an evaluation of task A.[*]

Task A|B - Article Violation given Allegation We further disentangle the LexGLUE tasks and pose *ECtHR A|B*. Given the facts of a case and the allegedly violated articles, the model should predict which (if any) specific articles have been violated.

[*]The LexGLUE dataset does not contain metadata (case id, Respondent state etc); in this work we use an enriched version of the same dataset by Mathurin Aché.

[*]The annotation explanations in (Chalkidis et al., 2021) state that "*The annotator selects the factual paragraphs that "clearly" indicate allegations for the selected article(s)*". We hypothesize that the so annotated passages contain information that is legally relevant for the violation as well.

This task reflects the legal process, as the court is aware of allegations made by the applicants when deciding. Providing information about the allegations shifts the nature of the task from topic classification to article-specific violation/non-violation prediction, thus refocusing the model and ideally leading to violation-specific explanations.

3.1 Data Distribution & Preprocessing

In order to facilitate model alignment, we worked with our ECtHR expert to identify shallow prediction signals in the fact statements that are unrelated to the legal merits of the complaint.

3.1.1 Length and Respondent State

For the task J dataset of Chalkidis et al. 2019, we find that the distribution of fact description length (number of sentences) and the distribution of respondent states are different between the two classes (see Appendix A). We hence account for the identity of the respondent state and the length of the fact descriptions via our deconfounding procedure for both datasets.

3.1.2 Accounting for Inadmissible Cases

We also observe in the task J dataset that the magnitudes of the running paragraph numbers differ between the classes, and that the single word "represented" strongly correlates with the positive class. This phenomenon arises because 2.6k of the 7k training cases are 'inadmissible' cases labeled as 'non-violation'. Legally, inadmissible cases are not necessarily 'non-violation' as inadmissibility relates to complaints not fulfilling the court's formal or procedural criteria.[*] In such cases, the court does not examine the merits of the application. The more interesting non-violation cases are such that are admissible, but in which no violation of the convention has been found. The single negative class contains instances of both inadmissible and admissible-but-no-violation-found cases. As explained above, the input texts of Chalkidis et al. 2019 are extracted from the FACTS section of full ECtHR decisions. In inadmissible cases, the applicant's background information can typically be found at the beginning of that section. We found that almost all inadmissible case facts start with the same formulaic sentence stating the applicant's name, nationality, and legal representation. This specific sentence is absent from the texts of admissible cases (violation and non-violation), where that information is part of a separate PROCEDURE section not included in the dataset. Moreover, due to the PROCEDURE section preceding the FACTS section in admissible cases, the running paragraph numbers appearing in FACTS sections of inadmissible cases are smaller than those of the admissible cases. If not remedied, these phenomena provide a considerable predictive signal for the label and distract the system from legally relevant information. In our experiments, we hence remove paragraph numbers from the input via preprocessing and account for distractor vocabulary via our deconfounding procedure described in Sec. 4. Still, the nature of task J remains unchanged and requires the system to classify the outcomes of a collection of both admissible and inadmissible cases.

3.1.3 Article-Specific Violation

By contrast, the more recent LexGLUE dataset only contains admissible cases and corresponding information about which articles the claimant has alleged to have been violated (for task B) along with those that the court has found to have been violated, if any (task A). The collection covers 10 different convention articles that make up the largest share of ECtHR jurisprudence. Each article has been alleged in a partition of the cases, and has been found to be violated in a subset of these.[*] For a given article in task B, all cases in which it has been alleged can be considered positive instances while the remaining cases are negatives. We consider task B as akin to topic classification, where the rights enshrined in the convention articles (e.g., Art. 6: right to a fair trial; Art. 1 Protocol 1: protection of property, etc.) may correlate with certain case fact language (e.g., related to law enforcement or expropriation, respectively). Task A incorporates this step and adds violation prediction per article, which is more difficult in principle. However, we observe that a few articles account for a large portion of the data and the conditional probability of a positive violation label in task A given its allegation labels from task B can be very high (see App. B). This makes an analysis of what trained models focus

[*] For example, the applicants lodge the complaint outside the time limit after the final domestic judicial decision or fail to exhaust required domestic remedies before complaining to the ECtHR, etc. It should be noted that the majority of inadmissible cases are decided by single judges and not available on the public database HUDOC.

[*] A few cases exist where the court refocuses the issues and finds a violation of an article that has not been alleged, but in the dataset they only occur in a negligibly small number of instances. (see App. Sec. B)

on more difficult, since they may learn to identify these dominant articles with high conditional violation probability, and be distracted from focusing on information that specifically signals violations of those articles. To remedy this, we propose task A|B that provides models an easy access to the label information of B, facilitating their focus only on determining whether the court finds a violation of given articles. This task is realistic since the allegations by the claimant are known to the court at the time that it decides whether the respondent state has violated the convention in the case.

4 Expert-Informed Deconfounding

We apply an expert-informed deconfounding method designed to mitigate the distracting effects of confounding elements and spurious correlations. As Pryzant et al. (2018) observe, accounting for confounders is common practice throughout many data analysis tasks to capture the intended signal and facilitate explainable models. In LJP, we understand confounding elements as such that influence both the observed legal outcome (convention violations found by the court as coded in the dataset) and the input text from which this outcome is to be predicted (here: ECtHR fact statements). Already covered examples are the different distribution of information across sections for admissible and inadmissible cases and the length of the fact descriptions (inadmissible cases tend to require less factual information to be dismissed).[*] An example of a spurious correlation is the identity of the respondent state (certain article violations will be claimed more often against a small number of governments, leading to a correlation). They each should have no bearing on the probability of an outcome in a given case as a judge will not decide against a violation because the facts are short, or because the case is against a particular government.

Confounding effects and spurious information in LJP may not be known ahead of time, especially if the legal decision is not made on the basis of an immutable a priori document, but rather on the basis of text that is technically a part of the eventual judgment. Our expert-informed method is intended to mitigate such situations where spurious correlations are introduced in the text production but may not be known in advance as explicit confounders.

Our method consists of two steps: (i) Identification of distracting attributes for deconfounding through a combination of simple model training and minimal expert markup, and (ii) mitigation of these effects through adversarial training.

4.1 Step 1: Identification of Distracting Attributes and Tokens

We first identify input attributes and categorize them as either distracting or genuinely legally relevant in an expert consultation. 'Distracting' attributes are highly correlated with the task label but not relevant in a human expert prediction. Attributes can be either (i) explicit in the text (such as vocabulary tokens) or (ii) implicit (e.g., country, text length, etc.). Implicit attributes can be derived from available metadata or a corpus analysis.

For textual attributes, we apply depth-limited decision trees on an n-gram representation of the fact statement to predict the case outcome. We extract all tokens that appear in the trees and iterate, successively removing tokens identified as predictive. Compared to extracting tokens from a single larger tree, this process is better suited to remove high-entropy-reducing tokens one typically finds near the root of trees. The list of removed tokens is then presented to a legal expert, who categorizes them into spurious and legally genuine (see Appendix Sec. F for the list of spurious vocabulary identified by the expert and the rationale behind the choices). This requires substantially less effort from the expert compared to other methods, such as data annotation or manual creation of counterfactuals. To prevent trees from picking up very sparse tokens, we filter the extracted terms using local mutual information (LMI) (Schuster et al., 2019), a re-weighted version of pointwise mutual information (PMI) (Church and Hanks, 1990). We calculate LMI for each pair of token and label as illustrated in Appendix Sec. G.

4.2 Step 2: Mitigation of Distracting Attributes

We assume a neural NLP model M consisting of a feature extractor F and classifier C with parameters θ_f and θ_c, respectively. For each confounder k, we apply a discriminator D_k with parameters θ_{d_K} to the feature extractors. We use adversarial training to maximize the feature extractor's ability to capture information for the main classification target while minimizing its ability to predict the

[*] If one assumes sectioning to be dependent on the case outcome, then this could even be characterized as an inverse causality relationship. For simplicity, and to account for court-internal document production processes, we understand "confounder" as including such configurations.

value of distractor attributes. This encourages the model to generate distractor-invariant feature representation for the classifier. We use the following adversarial training objective:

$$\sum_{k} \arg\min_{\theta_{d_k}} L(D_k(F(x)), y_k) \quad (1)$$

$$\arg\min_{\theta_f, \theta_c} \left[L(C(F(x)), y_c) - \sum_{k} \lambda_k L(D_k(F(X)), y_k) \right] \quad (2)$$

where L represents the loss, λ is a hyperparameter, x is the input, y_c is the label, and y_k is the distracting attribute k. The above optimization is performed using a gradient reversal layer (GRL) (Ganin and Lempitsky, 2015) to jointly optimize all the components instead of alternately updating the components as in GANs (Goodfellow et al., 2014). The GRL is inserted between the feature extractor and discriminators. It acts as the identity during the forward pass but, during the backward pass, scales the gradients flowing through by $-\lambda$, making the feature extractor receive the opposite gradients from the discriminator. This changes the overall objective function to :

$$\arg\min_{\theta_f, \theta_c, \theta_D} \left[L(C(F(x)), y_c) + \sum_{k} \lambda_k L(D_k(GRL((F(X))), y_k) \right] \quad (3)$$

We hypothesize that learning distractor-invariant feature representations through adversarial learning will help the model to focus on parts of the input that experts consider relevant.

5 Experiments & Discussion

In this section we describe our experiments in using our proposed deconfounding methodology to improve the alignment of model focus on the input with expert rationales on our set of LJP tasks.

5.1 Models

Baseline: We use the BERT variant of Hierarchical Attention Networks (Yang et al., 2016) as a baseline model. To segment our very long input texts we resort to a greedy sentence packing strategy in which we pack as many sentences as possible into one packet until it reaches the predefined maximum length (512 tokens constrained BERT). When a sentence exceeds this maximum, we split it into parts to fit into multiple packets. We encode each packet with LegalBERT (Chalkidis et al., 2020) to obtain the token level representations. Following Yang et al., 2016, we use a token attention layer aggregating the representation of the tokens and form a sentence (packet) vector. We pass these sentence vectors through a GRU encoder to obtain contextual representations. These are aggregated at the document level using a sentence attention layer. This model constitutes the feature extractor component F in our architecture. The obtained document representation is passed through dense layers for the final target prediction, constituting our classifier component C.

paraRem: Same as the baseline model but trained on data from which the paragraph number artifacts have been removed (see Sec. 3.1).

gradCou: *paraRem* model extended with a multi-class discriminator with a cross-entropy loss predicting the identity of the respondent government, and a corresponding deconfounding GRL.

gradLen: *paraRem* model extended with a length discriminator predicting the length (number of sentences) of the document via a set of bins and a cross-entropy loss, and a corresponding GRL to predict the bin value.

gradVocab: *paraRem* model extended with a vocabulary discriminator to predict the presence of identified spurious tokens, and associated GRL. As there can be multiple spurious tokens in a document, we employ binary entropy loss per token as it is a multi-label classification.

We refer to the above three deconfounded models collectively as *singleGrad* models.

gradAll: *paraRem* model extended with all country, length, and vocabulary discriminators in parallel, and associated GRLs.

Please refer to Appendix Sec. H for details on model configuration and training.

5.2 Quantitative Evaluation & Discussion

5.2.1 Expert Alignment Evaluation

Our main objective is to evaluate the alignment of the model's focus on the input text with legal expert rationales (i.e., selected subsets of relevant segments of the input). Following Chalkidis et al. 2021, we measure the model's ability to identify the correct rationales at the paragraph level, which is the natural granularity of ECtHR fact sections. To extract the importance score for each paragraph, we rely on an interpretability technique which quantifies the impact of a particular input token towards the final prediction of the model.

We use integrated gradients (Sundararajan et al., 2017) to obtain a token-level focus score and ag-

	J		B	A	A|B
Model	**valid**	**test**	**test**	**test**	**test**
Random	38.67 (4.52)	28.22 (3.16)	36.65 (2.91)	36.65 (2.91)	36.65 (2.91)
baseline	39.04 (4.31)	29.73 (3.28)	39.07 (2.94)	40.10 (3.02)	41.36 (3.02)
paraRem	41.81 (3.59)	31.53 (3.43)	41.47 (3.06)	41.93 (2.88)	41.86 (2.86)
gradCou	42.29 (3.54)	33.37 (3.75)	42.36 (3.09)	43.56 (2.78)	43.16 (2.85)
gradLen	42.21 (3.55)	33.58 (2.98)	*43.55* (2.92)	43.12 (3.31)	43.75 (3.14)
gradVocab	*42.96* (3.57)	*33.77* (3.41)	43.34 (3.02)	*44.48* (3.36)	*44.39* (2.86)
gradAll	**44.42** (3.40)	**34.48** (3.74)	**44.84** (3.13)	**45.95** (3.09)	**44.91** (3.18)
p-value	0.164	0.013	0.031	0.008	0.071

Table 1: Expert alignment performance expressed in precision@Oracle scores; value in brackets indicates standard error of the computed score; p values compare gradAll versus paraRem and were computed using a paired t-test.

gregate paragraph-level scores as the squared L2-norm of token scores in the paragraph divided by the square root of its number of tokens to account for length variation. We compute precision@k conditioned on some fixed k between the top-k paragraphs based on paragraph scores and golden paragraph rationales. The number of relevant paragraphs in gold rationales varies considerably, so a predefined k is inadequate. Thus, we compute precision@Oracle following Chalkidis et al., 2021, where *Oracle* is the number of relevant paragraphs in the gold rationales.

For tasks J, A, and A|B, the negative label (i.e., non-violation) is of similar interest as the positive label. In task B, however, the negative label merely indicates that a specific article has not been alleged, which is legally largely uninteresting. Hence, we reduce negative IG scores of tokens (indicating a negative contribution to the prediction) to zero.

5.2.2 Prediction Performance Evaluation

We also report the models' performance on the main four LJP tasks. For Task J, we report the macro F1-score for binary violation prediction. For Task A and B, following (Chalkidis et al., 2022a), we report micro-F1 (μ-F1) and macro-F1 (m-F1) scores. For Task A|B, we also report micro-F1 and macro-F1 scores. In computing the above metrics for tasks A and A|B, we consider the cases in which a particular article has been deemed violated as positive instances and the rest of the instances as negatives. We also introduce *hard-macro-F1* (hm-F1) for both Task A and A|B, in which F1 is computed for each article using only those instances as negatives where an article has been alleged as violated but not found so by the court.

5.2.3 Quantitative Evaluation Results

Table 1 and Table 2 show the performance of different models on expert alignment and outcome prediction, respectively.

paraRem vs. baseline: We observe that paraRem outperforms the baseline model in expert alignment across all tasks with a minimal drop in prediction performance. Task J stands out in that removing distracting signals via paragraph number removal even leads to a marginal improvement. Notably, we separately confirm the vulnerability of the baseline model by applying it to the test set with paragraph numbers removed and evaluate it on a test set without paragraph numbers, resulting in macro-F1 of 51.16 (i.e., a nearly 30 points drop).

gradCou, gradLen, gradVocab vs. paraRem: In all tasks, we observe that all singleGrad models improve in expert alignment performance over paraRem by a small but consistent margin. This demonstrates the ability of our deconfounding component to help the model better identify legally relevant parts of the input. Notably, gradVocab shows the most improvement in alignment over paraRem in all tasks except Task B (alleged article prediction), where gradLen performs best. During development on task B, we observed that the decision-tree based removal of predictive words led to only a marginal falloff in tree model accuracy, even after multiple iterations, since there was simply a lot of topical words (e.g., for police misconduct, legal proceedings, etc.) to take over as some of them were removed. This in part reflects the different nature of the tasks and shows a limitation of our tree-training-based method for identifying spurious tokens. Similar to paraRem, the gradLen model (in case of Task A, B, and A|B) also shows improvement in prediction performance compared to the baseline model. This suggests that deconfounding can potentially prevent the model getting stuck in distractor-related local optima.

Alignment: All singleGrad models outperform the baseline with regard to expert alignment. We observe that gradAll achieves the highest score, which establishes some degree of complementarity among the three singleGrad models and the distracting

signals they remedy. A paired t-test (gradAll vs. paraRem) reveals p-values above typical significance levels for the validation partition of task J, along with a considerable divergence in the general score level for the two tasks. We conjecture that this is the result of our small rationale sample size (50 from each partition) and differences in distribution between the task J data partitions, which have been split along the timeline rather than random. We also see a higher p-value for task A|B, which is intuitive since it is the most difficult. Its dataset lacks easily identifiable inadmissible cases (as in task J) and it has access to B's labels as concurrent, non-textual input. To gain some more insight into A|B, we report on a qualitative error analysis of the model rationales below.

5.3 Qualitative Evaluation & Discussion

Expert Scores: We sample 40 cases from task A|B validation and test sets (see App. Sec. D). We provide the expert with randomized visualizations of IG scores at the token level derived from our paraRem and gradAll models. Following (Jayaram and Allaway, 2021), the expert was asked to rate these on a five-point Likert scale (range -2 to 2) on two metrics: (i) Sufficiency: Is a sufficiently large set of tokens focused on to arrive at the prediction?; and (ii) Irrelevance: How many irrelevant tokens does the model focus on? We phrased the scale such that, for both parameters, a higher rating signals a better alignment between the model focus and the expert's assessment. Table 3 presents averages of the raw scale scores. We observe that the deconfounded gradAll model scores higher (See App. I for an example pair of IG visualizations).

Manual IG Inspection: For the paraRem model, we notice that high scoring IG tokens are sparse, whereas in gradAll, focus is densely distributed. There, contiguous spans of tokens tend to receive higher scores. This phenomenon is likely due to paraRem being drawn to single word distractors. Deconfounding helps the gradAll model to spread its focus across larger segments of the text. Our ECtHR expert further observed that gradAll highlighted words that, in conjunction, were indicative of the outcome, even if those were a considerable distance apart. At the same time, however, it seemed that two words hinting at opposite outcomes in a single sentence forced the system to focus only on one of the two, leaving the other one unhighlighted. We conjecture that these long- and short-distance phenomena are a result of the hierarchical model architecture necessitated by the long documents and leave their further exploration for future work.

An inspection of high scored tokens in paraRem reveals that many of them are highly discriminative in our decision tree models, showing that complex neural models can easily fall for distractors at the expense of missing equally predictive but semantically more complex signals. This reinforces our paradigm to identify discriminative tokens using a simpler model and subject them to expert scrutiny. In particular, we found that the word "represented" forms a natural decoy and, when injected into a violation-outcome fact statement, flips the predicted label of trained deep neural models. This led us to believe those models rely more on individual words than one might expect, and motivated us to explore how this can be exploited with information derived from simple models. Figure 2 shows that the performance of decision trees with unigram features (at iteration 1 without removed tokens) can even come close to BERT models.

In paraRem, we further observe that tokens at the start of sentences receive higher IG scores. We believe this to be the model counting sentences, which justifies deconfounding for length. For gradAll, we observe that sentence beginnings still receive focus, but less strongly so. This may be due to BERT recognizing sentence boundaries.

Further alignment improvement: The overall low precision@Oracle scores show that considerable differences in alignment with human experts remain. We conjecture that the model is shifting its focus, at least in part, to other spurious attributes which our current setup could not reveal. This calls for further investigation to design effective methods to identify such patterns. However, we expect them to be increasingly subtle and difficult to recognize, potentially even for legal experts. An intuitive upper bound for the system would be the annotation agreement of multiple experts, which to the best of our knowledge remains unexplored in the current state of the art.

Expert Pattern Identification: Our results naturally raise the question of how distractors can be identified in ECtHR fact texts by experts. Generally, the patterns we focused on affect the relationship between the argumentation in the judgment and the supportive facts given. There is copious literature on the court's inconsistent approach to

	Task J	Task B		Task A			Task A|B		
Model	**m-F1**	**μ F1**	**m-F1**	**μ F1**	**m-F1**	**hm-F1**	**μ F1**	**m-F1**	**hm-F1**
baseline	81.23	**78.08**	**68.42**	**69.28**	*58.80*	55.30	77.42	68.23	58.95
paraRem	**82.67**	77.82	66.90	68.94	58.35	55.62	77.20	67.32	58.80
gradCou	81.22	76.31	66.58	67.47	56.40	54.04	76.83	66.85	58.14
gradLen	81.99	*78.06*	*67.19*	*69.18*	**58.88**	**56.10**	**79.07**	**69.79**	**61.24**
gradVocab	*82.00*	77.68	66.40	69.06	58.47	*55.71*	*78.87*	*69.49*	*60.64*
gradAll	81.53	77.46	66.75	68.32	58.06	53.74	78.71	69.26	60.58

Table 2: Prediction Performance

Metric	paraRem	gradAll
Sufficiency (↑)	0.150	**0.550**
Irrelevance (↑)	0.475	**0.625**

Table 3: Qualitative evaluation scores

legal decision-making (e.g., Madsen et al. 2018) and it is known to switch between judicial policies depending on case circumstances (Helfer and Voeten, 2020). We hence paid attention to specific markers in the fact section and correlated them to existing precedents and argumentation patterns. A few examples: The court may decide to make use of positive obligations and decide against the state (violation) by highlighting failures of national authorities, or may decide to use those same positive obligations under 'the responsible authorities' doctrine, highlighting the efforts of national authorities to bring domestic legislation in line with the convention, thus deciding that there has been no violation. There are also fact patterns and practices specific to particular state parties to the convention (e.g., prison overcrowding, procedural issues in child abduction cases). The court may also sometimes highlight specific facts of a case with the view to 'document' its resemblance to, or divergence from, an existing precedent. A detailed, legally informed case study on predictive patterns is beyond the scope of this work.

5.4 Recommendations for LJP Research

In order to produce value for legal practice, we believe that LJP/LJF as an NLP task should strive for a productive combination of expert knowledge with data-derived insight. Based on our results, we formulate the following recommendations: First, as has already been observed in the field, any prediction/classification should happen from suitable source text that does not encode information about the outcome but contains as complete factual information as possible, or at least control for this influence. Second, straightforward predictors (e.g., input length and shallow unigram models) should be used to identify distractors and confounders. Third, claimed performance levels in predicting case outcomes should be contextualized by information about the distribution of the legal issues and respective conditional outcome probabilities in the corpus, as well as against baseline classifiers capable of exploiting known distractors. Fourth, more granular outcome variable information (e.g., case declared inadmissible vs. case dismissed on the merits, decomposition into outcomes of individual issues) will allow the development of more nuanced prediction/classification systems. Taken together, if such models can be explained and integrated into a decision support system for suitable tasks in legal practice, experts will be more likely to perceive them as adding value.

6 Conclusion

Our results show that our deconfounded LJP models are consistently better aligned with expert rationales than a baseline optimized for the target label only, and in many cases can even achieve better prediction performance. However, the improvement is small and the paragraphs focused on by all our models are still quite different from what an expert has annotated as relevant, as indicated by generally low precision@Oracle scores (<50%). Still, our quantitative results show that expert-informed deconfounding LJP works in principle and can potentially go a long way to train more robust and trustworthy neural LJP models, as well as derive more useful legal insight from them.

Limitations

We present a case study in deconfounding legal judgment prediction on the ECtHR, and all results are to be understood as relative to the ECtHR, its jurisprudence, the used datasets, and the formal tasks. The distracting attributes we identify include

confounding effects of the court's document production, where the decision may be known before the decision text (including the fact section) is finalized. A replication of this study in other LJP settings is of course warranted before general applicability can be claimed. Our analysis of task B has further revealed that redundant vocabulary distribution can challenge the system's ability to point out individual 'smoking gun' distracting tokens. This aspect is particularly complex in light of differing legal systems and their respective cultures and patterns of drafting texts that may form the basis of predictive or, more generally, assistive systems. Morphologically rich languages, where distracting signal may be spread across multiple tokens, may make this challenge more difficult and require stem- or lemma-based processing as part of the method.

Our deconfounding method is work-intensive and assumes the identifiability of distracting information in text and metadata by an expert. Legal expert agreement about what parts of decisions are relevant remains underexplored, and the division of genuine versus spurious language may also vary in between multiple experts. While we are convinced that further research on effective deconfounding of legal NLP systems is needed if these systems are to become robust and trustworthy, the time-intensive nature of collaboratively developing and qualitatively evaluating such models with legal experts poses a considerable resource challenge.

A technical difficulty in working with legal documents is their length, and the use of packet-based hierarchical models constrains the maximum distance across which tokens can directly attend to one another. The impact of this limitation on model performance in various types of tasks is the subject of ongoing exploratory work (e.g., Dai et al. 2022).

Ethics Statement

The research presented here works exclusively with publicly available datasets of ECtHR decisions, which are based on HUDOC[*], the public database of the Court. While these decisions are not anonymized and contain the real names of individuals involved, our work does not engage with the data in a way that we consider harmful beyond this availability.

Our models are designed to be used with pretrained language models and hence inherit any biases they may contain. This entails an obligation to screen incorporated models and to test any developed system with regard to its performance across groups of cases (e.g. Chalkidis et al. 2022b), and to remedy any disparities before deploying it as a prediction and inference tool. Our experiments are targeted at controlling for legally irrelevant distractors in the input, which is in line with this responsibility.

The task of legal judgment prediction raises ethical concerns, both general as well as specific to the European Court of Human Rights. (Fikfak, 2021) emphasizes focal issues with regard to the court considering the use predictive technology to tackle its caseload, including system bias and the challenges of designing the interaction between judges and predictive systems. The latter is of course especially sensitive given experiences made with recidivism risk prediction (Collins 2018) and possible disparate effects of how judges interact with scores (Albright 2019). Our research group is committed to research on LJP as a means to derive insight from legal decision data towards increasing accountability, fairness, and transparency in the use of technology in legal systems. The premise of this work is that the behavior of legal outcome prediction systems is to be scrutinized with great care. This paper does not advocate for the practical use of such systems, but rather empirically explores difficulties that arise in their development and recommends a closer connection between technical research and legal expertise (see Sec. 5.4).

All models of this project were developed and trained on Google Colab. Our models adapted pretrained language models and we did not engage in any training of such large models from scratch. We did not track computation hours.

Acknowledgments

We are grateful to Jaromir Savelka for the ability to use the *Gloss* annotation tool and for providing feedback on the draft. We also thank the anonymous reviewers for valuable comments.

References

Alex Albright. 2019. If you give a judge a risk score: evidence from kentucky bail decisions. *Harvard John M. Olin Fellow's Discussion Paper*, 85:16.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of

[*] https://hudoc.echr.coe.int

human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2021. Investigating the state-of-the-art performance and explainability of legal judgment prediction. In *Canadian Conference on AI*.

L Karl Branting. 2013. *Reasoning with rules and precedents: a computational model of legal analysis*. Springer Science & Business Media.

L Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2021. Scalable and explainable legal prediction. *Artificial Intelligence and Law*, 29(2):213–238.

Stefanie Brüninghaus and Kevin D Ashley. 2005. Generating legal arguments and predictions from case texts. In *Proceedings of the 10th international conference on Artificial intelligence and law*, pages 65–74.

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, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022a. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022b. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Erin Collins. 2018. Punishing risk. *Geo. LJ*, 107:57.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. *arXiv preprint arXiv:2204.06683*.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Ayse B Dincer, Joseph D Janizek, and Su-In Lee. 2020. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics*, 36(Supplement_2):i573–i582.

Veronika Fikfak. 2021. What future for human rights? decision-making by algorithm. *Decision-making by algorithm (September 3, 2021). Strasbourg Observers*, 19.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4957.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Matthias Grabmair. 2017. Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs. In *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law*, pages 89–98.

Laurence R Helfer and Erik Voeten. 2020. Walking back human rights in europe? *European Journal of International Law*, 31(3):797–827.

Sahil Jayaram and Emily Allaway. 2021. Human rationales as attribution priors for explainable stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5540–5554.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in nlp. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691.

Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2019. Legal cause prediction with inner descriptions and outer hierarchies. In *China National Conference on Chinese Computational Linguistics*, pages 573–586. Springer.

Luyao Ma, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Wei Ye, Changlong Sun, and Shikun Zhang. 2021. Legal judgment prediction with multi-stage case representation 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.

Mikael Rask Madsen, Pola Cebulak, and Micha Wiebusch. 2018. Backlash against international courts: explaining the forms and patterns of resistance to international courts. *International Journal of Law in Context*, 14(2):197–220.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Masha Medvedeva, Ahmet Üstün, Xiao Xu, Michel Vols, and Martijn Wieling. 2021. Automatic judgement forecasting for pending applications of the european court of human rights. In *ASAIL/LegalAIIA@ ICAIL*.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28(2):237–266.

Masha Medvedeva, Martijn Wieling, and Michel Vols. 2022. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, pages 1–18.

Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. 2018. Deconfounded lexicon induction for interpretable social science. 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 1615–1625, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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 3419–3425.

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.

Mehmet Fatih Sert, Engin Yıldırım, and İrfan Haşlak. 2021. Using artificial intelligence to predict decisions of the turkish constitutional court. *Social Science Computer Review*, page 08944393211010398.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3086–3095, Online. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

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.

A Dataset Statistics

Table 5 demonstrates the artefacts from corpus construction and admissibility-related confounding information in the training set of Task J. Figure 3 and 4 display the distribution of text length and the respondent state in the Task J train set. Figure 5 and 6 show the statistics of text length and respondent state in the Task B train set.

B LexGlue Dataset Characteristics

Table 4 describes the conditional probability of a violation finding by the court given the allegation of a particular article as well as the probability of a violation finding regarding a particular article even though it was not alleged.

C Rational annotation Process for Task J

We sampled 50 cases (25 each) from the validation and test split. In each split, we sample two cases for each of the ten violated articles, one containing the token 'represented' and one without, along with five inadmissible cases. While the article information is available in the task J dataset, we do not use it as it was introduced as a binary violation classification task.

The rationale annotation process was done using the *GLOSS* annotation tool. The third author of this paper, who is an ECtHR expert, read the case fact statements and highlighted paragraphs which she considered indicative of an eventual finding of a violation for any convention article by the court. Despite our sampling involving randomness, the expert was already familiar with a considerable portion of the decisions. Given this, we abstained from producing a human expert outcome prediction baseline.

D Case Sampling for Qualitative Evaluation

For the qualitative evaluation of Task A|B, we sample 40 cases (20 each) from validation and test split. In each split, we sample two cases for each of the ten *allegedly violated* articles, one with a finding of a convention violation and with a non-violation finding.

E Decision Tree Performance

Figure 2 shows the performance of our decision tree model across iterations for different tasks. After each iteration, we remove the informative to-

kens from previous iterations. In case of task J, we notice a steep fall after iteration 5. Tasks A and B exhibit less dramatic falloff of macro-averaged F1, owing to the different nature of the tasks as article-specific violations. Performance on Task A|B even shows small increases, albeit with a low absolute score. The large standard deviations bands computed across all articles show considerable variation.

F Spurious Vocabulary identified by Expert

Following is the spurious vocabulary we obtained with respect to each task.

- **Task J:** represented, national, mr, summarised, practising, lawyer, agent, paragraph
- **Task B:** hearing, born, adjourned, detained, noted, hearing, alleged, investigation, place, question, members
- **Task A:** stated, could, also, one, arrested, detained, hearing, investigation, hearing, within, due, second, hearing, certain
- **Task A|B:** february, published, november, march, religious, investigation, first, service, letter, carried, would, one, submitted, head, march, damage, group,provided, seen

The words were chosen as relevant or irrelevant by using the daily vocabulary of a human rights lawyer working at the ECtHR as a reference. A word was considered legally relevant if, taken individually, it could be introduced into legal reasoning. For instance, the word "religious" was spurious because taken individually it says nothing about the content of a norm. One may talk about religious freedom, but the legally relevant word there is freedom. Article 9 mentions religion, but restrictions related to religion may also be present under Article 8, 3, 2, 5, etc. Under the same Article 9 for instance, the court decides whether there has been a violation depending on criteria such as tolerance, pluralism, etc. It is those criteria that are relevant whereas "religion" is not by itself relevant as a part of the legal reasoning.

G LMI Calculation

We calculate LMI for each pair of token t and label y as follows:

$$LMI(\text{t, y}) = p(\text{t, y}) \times PMI(\text{t, y}) \quad (4)$$

$$p(\text{t, y}) = \frac{\text{count(t,y)}}{|D|} \quad (5)$$

$$PMI(\text{t, y}) = \log \frac{p(\text{t | y})}{p(\text{t})} \quad (6)$$

where $count(t, y)$ denotes the co-occurrence of t and label y, and $|D|$ is the number of unique words in the training set.

In the case of binary classification (task J) and one-vs-one multi-label classification (task A|B), we calculate the LMI score for a token as the absolute difference between LMI scores for both positive and negative labels, as both the labels represent a particular class. In one-vs-rest (tasks A, B), we simply take the difference between LMI scores for both positive and negative labels (rather than absolute difference) as the negative label does not specifically represent a particular class. Finally, we calculate the z-score statistic of the effective LMI score for each token to identify significant tokens.

H Model configuration & Training

Spurious token identification: We train a series of decision trees of depth 3 to assemble lists of predictive tokens for expert filtering. The feature vector consists of whitespace-tokenized unigrams reduced by the LMI filtering explained above. For task J, this means training trees that predict the binary violation label. For task A and B we employ a one-vs-rest classification to produce one decision tree series per article. For task A|B we provided the task B labels (allegedly violated articles) in one-vs-one fashion per article, with positive instances being facts where that particular article was deemed violated, and negatives where that particular article was merely alleged but not deemed violated.

LJP models: Our models compute BERT-based word embeddings of size 768. Our word level attention context vector size is 300. The sentence level GRU encoder dimension is 200, thus giving a bidirectional embedding of size 400, and a sentence level attention vector dimension of 200. The final dense classifier for all tasks has 100 hidden units. The output dimension is 1 for task J and 10 for the other tasks (i.e. one per convention article). For task A|B, we concatenate a multi-hot 10-element feature vector containing the task B labels to the output of the feature extractor before it is passed to the classifier. All discriminators (country, length, and vocabulary) are built as analogous classifiers with a hidden dimension of 100 and output layer

dimensions as required by each of them. We use mini batches size of 8 in case of Task J and 16 for all other tasks. The model is optimized end-to-end using Adam (Kingma and Ba, 2015). The dropout rate (Srivastava et al., 2014) in all layers is 0.1. We determine the best learning rate using a grid search on the development set and use early stopping based on the development set F1 score.

I Visualization of IG score

Figure 7 exhibits screenshot excerpts of a sample case text provided to the legal expert for qualitative evaluation. The yellow background highlight was not in the original visualization and has been supplied here as a reference. We add it here as an example of focus patterns shifting incurred by our deconfounding method.

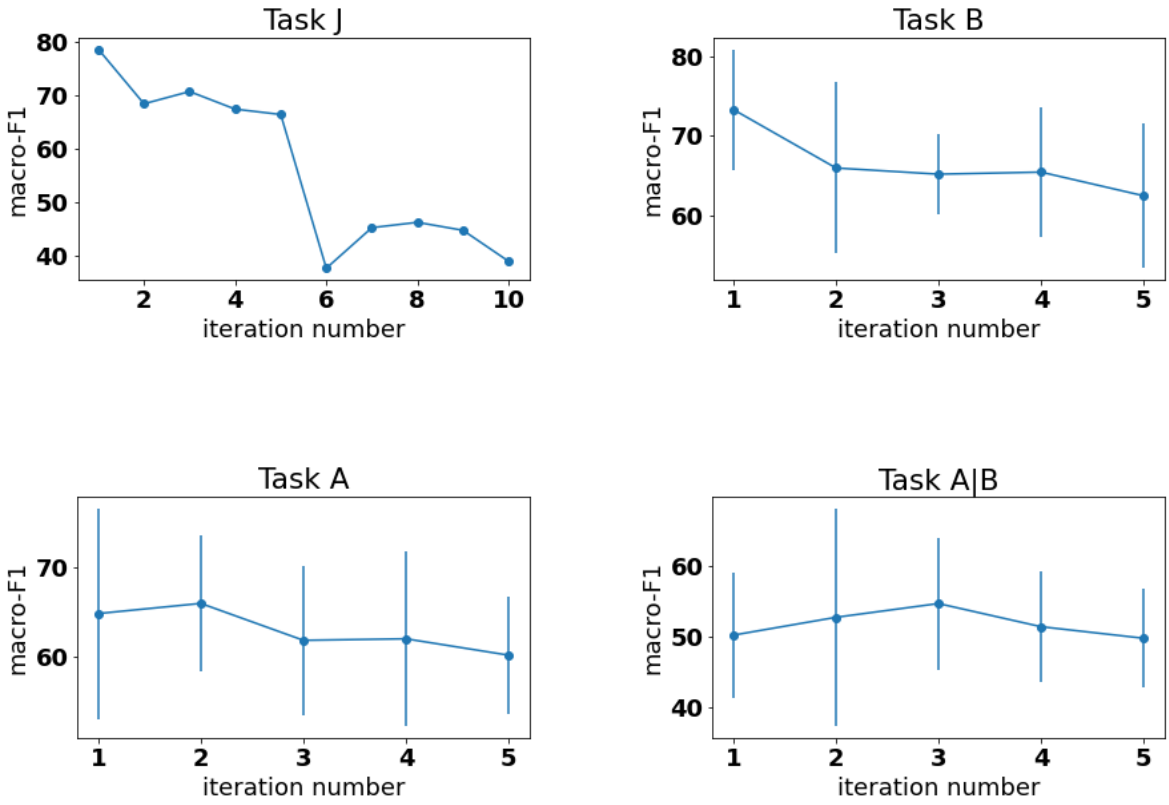


Figure 2: Macro-F1-Performance of Decision Trees across different iterations with removal of informative feature nodes in successive iterations. For Task A,B, A|B, standard deviation bars represent variability in F1-score across the 10 articles.

Article	**% of cases alleged**	**% of cases alleged and violated**	**% of cases alleged but not violated**	**conditional probability of violation given allegation**	**% of cases not alleged but violated**
10	4.9	3.1	1.8	63.27	0.13
11	1.8	1.2	0.6	66.67	0.02
14	4.93	1.51	3.42	30.63	0.06
2	6.92	5.48	1.44	79.13	0.13
3	19.33	14.5	4.83	75.0	0.49
5	18.03	14.68	3.36	81.39	0.52
6	60.41	51.64	8.77	85.49	0.62
8	11.73	7.66	4.08	65.25	0.23
9	0.9	0.44	0.46	49.38	0.01
P1-1	17.31	15.02	2.29	86.78	0.77

Table 4: Statistics of LexGlue Train Dataset (Total of 9000 Cases) (Chalkidis et al., 2022a)

	violation	non-violation
number of cases	3551	3549
avg. number of sentences per case	55	35
number of inadmissable cases	0	2608
avg. first paragraph number appeared in the text	6.1	1.8
percentage of cases containing the word 'represented'	0.17	0.68

Table 5: Statistics of violation (label 1) and non-violation (label 0) cases in the training set of Task J. Some inadmissible cases are so short that no paragraph numbers appear. In such situations we count the paragraph number as 0.

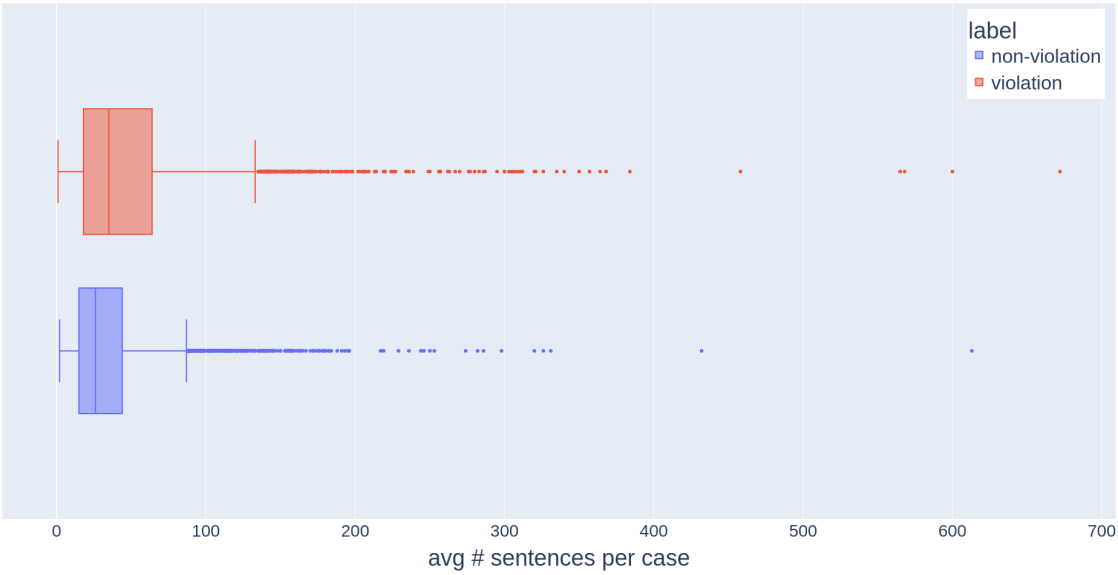


Figure 3: Text length distribution in the training set of Task J

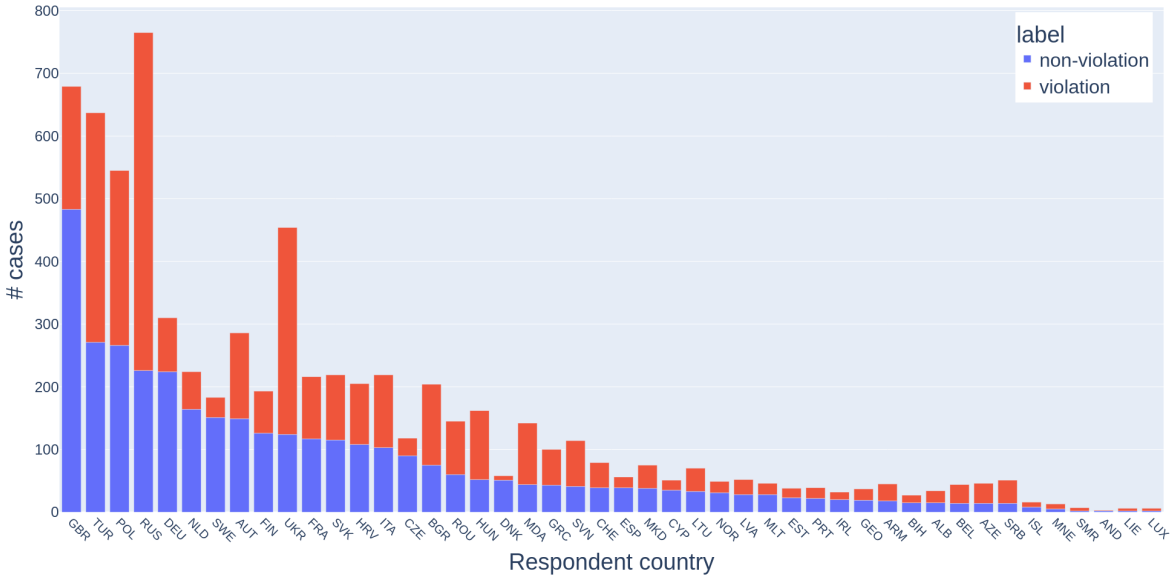


Figure 4: Country distribution in the training set of Task J

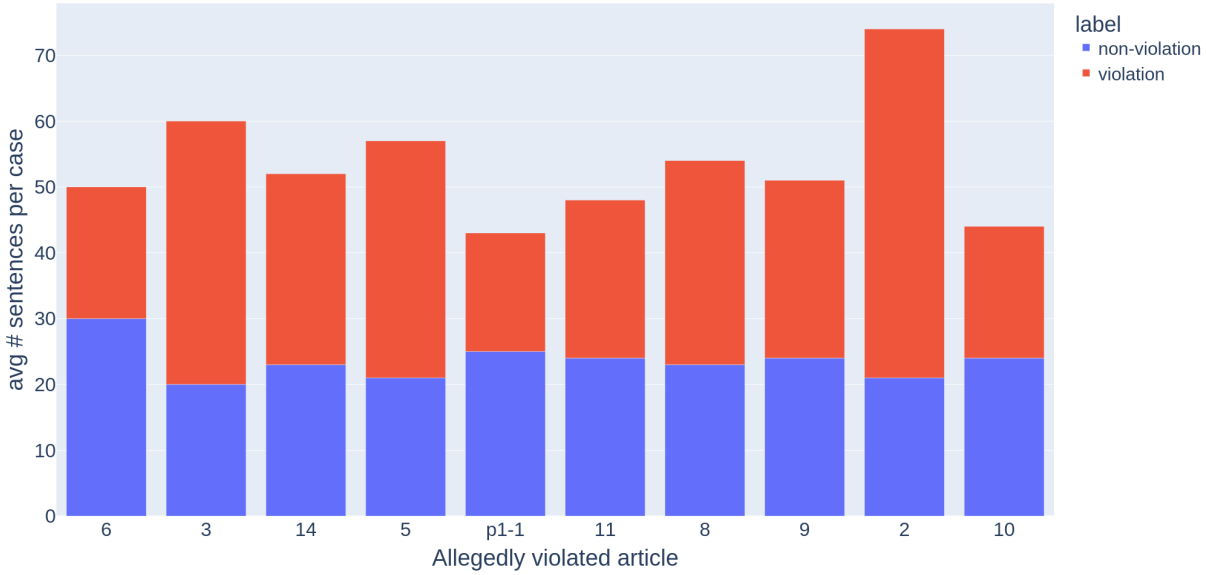
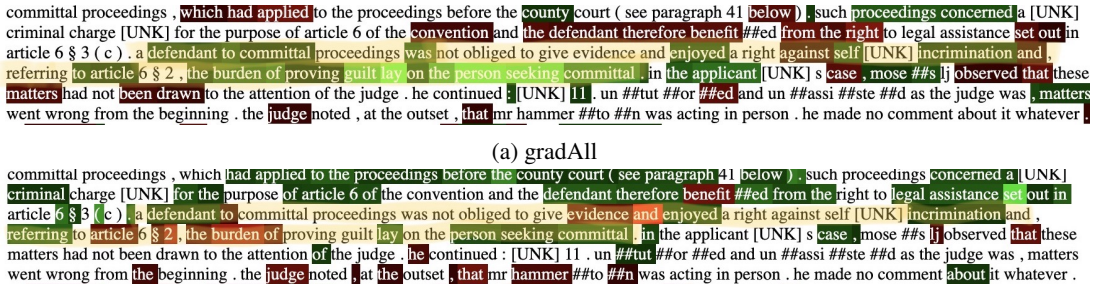


Figure 5: Text length distribution for each violated article in the training set of Task B

Figure 6: Country distribution in training set of Task B. For the display effect, only countries accounting for more than 1% of the total cases are displayed.



(a) gradAll

(b) paraRem

Figure 7: Example visualizations of different IG scores derived from the (a) gradAll and (b) paraRem model, respectively. The gradAll model focuses contiguously and densely on the expert-annotated indicative sentence 'A defendant to committal proceedings ... proving guilt lay on the person seeking committal' (yellow background highlighted); while the paraRem model fails to focus the latter half the indicative sentence.