Comparative Study of Explainability Methods for Legal Outcome Prediction

Ieva Raminta Staliūnaitė
University of Cambridge
irs38@cam.ac.uk

Josef Valvoda
University of Copenhagen
jval@di.ku.dk

Ken Satoh
National Institute of Informatics
ksatoh@nii.ac.jp

Abstract

This paper investigates explainability in Natural Legal Language Processing (NLLP). We study the task of legal outcome prediction of the European Court of Human Rights cases in a ternary classification setup, where a language model is fine-tuned to predict whether an article has been claimed and violated (positive outcome), claimed but not violated (negative outcome) or not claimed at all (null outcome). Specifically, we experiment with three popular NLP explainability methods. Correlating the attribution scores of input-level methods (Integrated Gradients and Contrastive Explanations) with rationales from court rulings, we show that the correlations are very weak, with absolute values of Spearman and Kendall correlation coefficients ranging between 0.003 and 0.094. Furthermore, we use a concept-level interpretability method (Concept Erasure) with human expert annotations of legal reasoning, to show that obscuring legal concepts from the model representation has an insignificant effect on model performance (at most a decline of 0.26 F1). Therefore, our results indicate that automated legal outcome prediction models are not reliably grounded in legal reasoning.[1]

1 Introduction

Interpretability is at the core of legal practice. Lawyers and judges pour over legal text to interpret it in light of current affairs, the case at hand and the general zeitgeist (Valvoda et al., 2024). In the context of natural legal language processing, interpretability is no less important. Primarily, this is because the use of Machine Learning (ML) in law can have profound effects on human life (Hacker et al., 2020). This risk is widely acknowledged, as reflected in the EU's General Data Protection Regulation (GDPR), which mandates that legal decisions must be explainable (Hamon et al., 2020; Selbst and Powles, 2017).[2] As such, we advocate for interpretability to be a central focus of NLLP research.

Despite early contributions to the field, which include symbolic methods (Ashley, 1991; Collenette et al., 2020) as well as attention-based interpretability (Branting et al., 2021), and the emergence of recent domain-specific methods (Valvoda and Cotterell, 2024), there remains a lack of a comprehensive overview of the popular NLP interpretability tools that can be applied to legal contexts.

In this work, we offer such a comparative study. We focus on explainability of neural models in the context of the legal outcome prediction task - a popular NLLP task (Brüninghaus and Ashley, 2006; Zhong et al., 2018; Chalkidis et al., 2019; Long et al., 2019; Dong and Niu, 2021; Ma et al., 2021). In particular, we work on the recent reformulation of this task as a three-way classification (Valvoda et al., 2023) and compare three influential interpretability methods from general NLP on legal outcome prediction - Integrated Gradients (Sundararajan et al., 2017), Contrastive Explanations (Jacovi et al., 2021) and Concept Errasure (Ravfogel et al., 2022).

We first hypothesize that while different explainability methods might provide varying results in terms of what legal outcome prediction models use in their decision-making process, the models are likely to be using features which differ from those that a human would deem important. We put this hypothesis to the test by correlating the attributions from the methods of Integrated Gradients (Sundararajan et al., 2017) and Contrastive Explanations (Jacovi et al., 2021) with ground truth data from court rulings. Confirming our hypothesis, we measure a very weak correlation between the predicted importance scores and the ground truth labels, with the absolute values of the correla-

[1] Our code: https://github.com/ieva-raminta/XNLLP

[2] Specifically, Article 22 and provisions of Articles 13-15 of GDPR ask for a *'meaningful information about the logic involved'*.

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 243–258

tion coefficients of Spearman (1904) and Kendall (1938) ranging between 0.003 and 0.094.

We further hypothesise that the outcome prediction models are not likely to perform complex legal reasoning, such as those captured in annotated datasets of legal arguments (Habernal et al., 2023) and concepts (Mumford et al., 2023). We test this by using the Concept Erasure method (Ravfogel et al., 2022). Indeed, we find that the outcome prediction models perform on par or at times even better when legal concepts are obscured from their representations, with F1 on the outcome prediction task decreasing by at most 0.26 - a statistically insignificant change.

We conclude that the subpar performance of neural models on negative outcome prediction task is symptomatic of a larger issue - the models do not reason like a human legal professional would.

2 Related Work

Over the years, different approaches have been proposed to address the question of how an ML model *reasons* (Ribeiro et al., 2016; Lundberg and Lee, 2017). Researchers have compared the faithfulness and cost of various interpretability methods (Lipton, 2018; Wiegreffe and Pinter, 2019; Jain and Wallace, 2019; Wang et al., 2020). The majority of these explainability methods link surface features in the model input to model predictions. Some research has emphasized that there is likely no single feature-based explanation for a given model prediction (Camburu, 2020). Thus, the development of alternative, concept-based methods has complemented their feature-based counterparts (Yeh et al., 2020). Furthermore, humans find the explanations of deceptive machine learning systems equally convincing as those of truthful models (Pataranutaporn et al., 2021), which stresses the importance of explanations being faithful as opposed to simply convincing (Alhindi et al., 2018; Piratla et al., 2023; Atanasova et al., 2023).

The earliest work in explainability in NLLP are the legal reasoning systems of HYPO (Ashley, 1991) and CATO (Aleven, 1997). These symbolic systems involve a manual extraction of factors that do not deterministically influence the case outcome, but rather weigh the decision positively or negatively with varying strength, depending on the context. Since then researchers have developed hybrid systems, using stochastic methods to extract the features that are then fed into a rule-based system. Falakmasir and Ashley (2017) have used tf-idf and Latent Dirichlet Allocation (LDA), while Mumford et al. (2023) employed transformer models (Vaswani et al., 2017), to extract factors to be used in rule-based systems.

Researchers have also studied explainability in fully probabilistic methods. Branting et al. (2019) and Branting et al. (2021) use attention as explanation. Yamada et al. (2024) solve the outcome prediction and rationale extraction tasks via a multi-task approach. Norkute et al. (2021) evaluate the usefulness of attention scores as well as scores from a source attribution method based on word overlap, by measuring the increase in the speed of humans reviewing legal summaries. Strickson and De La Iglesia (2020) and Soh Tsin Howe (2024) use topic models along with other feature-extraction methods. Gray et al. (2023), Gray et al. (2024) and Drápal et al. (2023) use LLMs to extract factors in legal cases. Valvoda and Cotterell (2024) explore a novel interpretability method in NLLP, namely influence functions (Koh and Liang, 2017), in order to determine which cases in the training data influence the outcome predictions.

Some researchers have also addressed the problem that interpretability methods are difficult to evaluate, given that even humans disagree on what a correct explanation is. Malik et al. (2021) compare the results of an outcome prediction model with some parts of the input masked and without, which highlight the difference between the importance attributed by experts and the occlusion method. Salaün et al. (2022) have shown low agreement between integrated gradients scores of models and expert annotations. Feng et al. (2022) link the errors of outcome prediction models to their failure to detect the parts of the input that determine the judgment. Santosh et al. (2022) use deconfounding to align model predictions with expert reasoning. Xu et al. (2023) study human label variation with regard to the rationales explaining legal outcomes, and show low agreement not only between experts and models, but also among experts, which highlights the difficulty of the task. By and large, the models described in this subsection do not compare multiple interpretability methods.

Legal outcome prediction is the task of predicting the outcome of a case, i.e. whether a law has been violated, given the case facts, which describe the circumstances of the parties involved.[3] Over

[3] See appendix A for a condensed example of a case.

the years, predicting the outcome of a court case has been approached by many researchers in a number of jurisdictions (Virtucio et al., 2018; Chalkidis et al., 2019; Mumcuoğlu et al., 2021; Jacob de Menezes-Neto and Clementino, 2022; Cui et al., 2023). Perhaps due to its conceptual simplicity, the task is one of the cornerstones of NLLP research and is usually defined as a binary classification task (Feng et al., 2022; Cui et al., 2023). Different approaches to legal outcome prediction use the case facts (Shaikh et al., 2020), the complaints (Chalkidis et al., 2019; Semo et al., 2022), the contents of the laws (Zhong et al., 2018), and/or the facts of precedent cases (Cao et al., 2024).

Recent work has re-framed outcome prediction to reflect the reality of a court setting better (Valvoda et al., 2023). Instead of predicting if an article is violated or not, the task is to predict whether the article is claimed to be violated (simulating the role of a lawyer), and then whether it is actually found to be violated or not (simulating the role of a judge). This can be simplified as a three-way classification objective. In practice, a model is trained to predict **positive outcomes**, i.e. when a law has been claimed as violated and found as violated, **negative outcomes**, i.e. when a law has been claimed as violated but the judge found it was not violated, **null outcomes**, i.e. when a law has not even been claimed as violated and is irrelevant to the case.

3 Explainability for Models of Legal Outcome

In our explainability experiments, we focus on the re-framing of the legal outcome prediction task following Valvoda et al. (2023). We do this for three reasons. (1) The new formulation better reflects the actual legal process and outperforms prior work in the domain of the European Convention of Human Rights (ECHR). (2) legal outcome prediction models turn out to perform particularly poorly when having to predict negative outcomes. This is diametrically opposite to their excellence at predicting positive outcomes. Thus, at the core of our paper lies a natural question arising from this asymmetry. Why do the models struggle with negative outcome prediction? (3) Having three target classes instead of two, opens up new possibilities in terms of the explainability methods we can study.

We begin our work with the standard Explainable Artificial Intelligence (XAI) method - namely integrated gradients (Sundararajan et al., 2017). This method highlights the input tokens (in our setting facts) the model finds important for a given case. Then, we move to a contrastive explanation method (Jacovi et al., 2021). Since we work with a three-way classification problem we can employ contrastive explanations to infer which facts are particularly important to the model for distinguishing each target class from every other class. Our main interest here is to understand why the models struggle with negative outcomes.

Finally, we use the Concept Erasure method (Ravfogel et al., 2022) to perform a deeper analysis of whether any legal concepts are used by the model. Unlike the prior two methods where we study the effect of input tokens, here we study the effect of legal concepts encoded in the latent representations learned by the model. We describe each of the above approaches in more detail in section 5.

Given the ground-truth data of human annotated token sequences and legal concepts (Chalkidis et al., 2021; Habernal et al., 2023; Mumford et al., 2023), we can begin to study how well a model aligns with human judgement when it comes to legal reasoning. Furthermore, the chosen set of XAI approaches allows us to study the difference between superficial textual explanations versus explanations through concepts in the legal AI domain.

4 Parametrizing Legal Outcome Prediction Models

We finetune a sequence classification model on the ECtHR dataset to jointly predict the positive, negative and null outcomes, following the architecture of the ternary prediction setup from Valvoda et al. (2023). One modification made to the model architecture is replacing the multi-layer perceptron (MLP) with a simple linear classification layer. This change ensures that the setup is compatible with the interpretability methods discussed in Section 5, where linear classifiers are used.

We choose the LEGAL-BERT model (Chalkidis et al., 2020) due to its domain-specific training set and the fact that it yields the best performance for the ternary setup we are using for our experiments.[4] The model is trained on a single NVIDIA TU102 GPU with batch size 16, for a maximum of 10

[4]We replicate the experiments with other models too in order to ensure that the model results are not idiosyncratic to our chosen setup. Please see appendix C.

epochs, using early stopping by monitoring the loss.

The model performs the best on the null class, and yields particularly poor results on the negative cases, as shown in Table 1. The most common mistakes of the model are assigning the null label to the items from the positive and negative classes, as shown in Table 2.

Metric	null	positive	negative
precision	93.55	78.80	48.07
recall	98.68	77.93	10.33
F1	96.04	78.36	17.01

Table 1: Results of the three way outcome prediction LEGAL-BERT model on the ECtHR test set.

	True \ Pred.	null	positive	negative
=	null	12117	109	53
	positive	258	1056	41
	negative	580	175	87

Table 2: Confusion matrix of the three way outcome prediction LEGAL-BERT model on the ECtHR test set.

5 Interpretability Methods

This section describes the interpretability methods used in this study, along with their implementation.

5.1 Integrated Gradients

The Integrated Gradients method, or Axiomatic Attribution for Deep Networks (Sundararajan et al., 2017) is a gradient-based attribution method, which does not require any instrumentation of the network that it is being applied to. The model uses a baseline input as a counterfactual to each feature being tested for attribution. In the context of language, a sequence of the <PAD> tokens can be used for this purpose. Integrated gradients are obtained by accumulating the gradients collected along a path from the baseline to the input. In this work we use the implementation of Layer Integrated Gradients from the Captum package (Kokhlikyan et al., 2020), where we compute the attributions with regard to the BertEmbedding layer. This approach is chosen as a reliable yet simple interpretability method.

5.2 Contrastive Explanations

Jacovi et al. (2021) propose a Contrastive Explanations method for model interpretability that is inspired by cognitive science research. Since humans generate explanations contrastively, namely explaining why a certain occurrence happened instead of some alternative, they argue that XAI methods should mimic this type of reasoning. Hence, instead of comparing the input to a neutral input such as the baseline in the Integrated Gradients method, Jacovi et al. (2021) project the input representation onto a space which minimally separates two class labels as predicted by the model. The predicted label is called 'fact', and the alternative label 'foil'. The contrastive explanation can then be generated by computing the difference between the original representation and the contrastive projection. The method is also applicable to any neural classifier. The application of the contrastive approach is particularly interesting with the negative cases, given that they meaningfully contrast to the positive cases by virtue of not violating a given article, while contrasting to the null cases by allegedly violating the article.

5.3 Concept Erasure

A deeper interpretability method that we employ in this study, is at the level of concepts instead of surface level input features. Inspired by the idea of Jacovi et al. (2021) to use concept attribution for explainability, we apply Linear Adversarial Concept Erasure, presented by Ravfogel et al. (2022), to our task. That is, the method obscures concepts which may or may not influence the model predictions for the main task by projecting them to a space where a linear classifier can no longer recover the signal to determine the presence of the concept in the input. An adversarial model is trained with a constrained, linear minimax game to erase the concept while maintaining the performance on the main task. We interpret the outputs of this method to show the importance of a given concept to a trained model through the difference in model performance when the concept is erased from the input representation. The purpose of using this explainability method for the legal outcome prediction task is to investigate the use of actual legal concepts rather than relying on superficial input features.

In this study we adapt the Concept Erasure model to our multi-output concept prediction task, where for each concept the model solves a binary classification problem of predicting the presence of the given concept in the input document. We train a Logistic Regression model on the subset of the ECtHR training set which contains items annotated

for the presence of legal concepts. We use a maximum of 4000 iterations, l2 penalty, saga solver and warm start. The input to the classifier is the encoded representation from the last hidden state of the trained LEGAL-BERT model described in Section 4.

6 Datasets

We use the European Court of Human Rights (ECtHR) dataset (Chalkidis et al., 2021)[5] and its extensions for the experiments in this study (see Table 3 for data statistics). We are using the version of the dataset presented by Valvoda et al. (2023), as it is more complete. Valvoda et al. (2023) also extend the task of outcome prediction to include prediction of negative outcomes. The *Allegedly Violated Articles* and *Violated Articles* together comprise information about *Positive* cases and *Negative* cases. Namely, if an article is both *Allegedly Violated* and *Violated*, the case is positive, whereas if an article is only *Allegedly Violated* but not *Violated*, then the case is negative. The Silver Allegation Rationales indicate the parts of the case facts which are referenced in the decision of the judge. These rationales are available for all the cases where a regular expression match is found between the judgment and the case facts. Similarly, Gold Allegation Rationales have been annotated by a legal expert as the important facts for the allegations. Only 50 cases have been annotated with gold rationales. The silver rationales are more abundant, but less reliable than the gold ones. In this study we are only using the Silver Rationales, due to the size of the annotated data.

In addition, we are also using annotations of legal concepts in ECtHR. Firstly, Habernal et al. (2023) annotate a corpus of 373 court decisions covering Articles 3, 7, and 8, with legal arguments being made in each case. The purpose of the dataset is to aid Legal NLP models in coming closer to legal reasoning in modeling outcomes. Secondly, Mumford et al. (2023) annotate 735 cases pertaining to Article 6, with legal concepts that correspond to factors in a rule-based legal reasoning system. The concepts used in this study are listed in appendix B.

[5] https://huggingface.co/datasets/AUEB-NLP/ecthr_cases

6.1 Dataset Preprocessing

Integrated Gradients and Contrastive Explanations. We use Silver Rationales annotations as the target labels for evaluating the interpretability methods. We adjust the level of granularity of the outputs from the Integrated Gradients and Contrastive Explanation methods in order to make them comparable. The token-wise attributions from the Integrated Gradient method are accumulated per paragraph in order to match the paragraph-wise ground truth in Silver Rationales. Similarly, due to the cost of masking every token in very lengthy case fact documents, when applying the Contrastive Explanations method to the ECtHR data, we modify the masking method to cover entire paragraphs rather than single tokens. When a paragraph is being masked, it is replaced by a sequence of <MASK> tokens of a length equal to the number of tokens in that paragraph.

Concept Erasure. The Legal Argument (Habernal et al., 2023) and Legal Reasoning (Mumford et al., 2023) labels are transformed to binary targets. That is, we convert the token-wise sequence tags indicating the presence of legal arguments from the Legal Argument dataset to a binary document-wise label, indicating whether the concept is present in the case. From the available annotations in Habernal et al. (2023) we use the Argument Type data as the legal concepts. Similarly, the mean annotator scores for the presence of concepts in the Legal Reasoning dataset are converted to a binary label using the ARGMAX of [positive ascription annotations, negative ascription annotations, no ascription annotations] scores, and interpreting both positive ascription and negative ascription to indicate a presence of that concept. For both datasets, we only use a concept if it appears in at least one but not all of the cases in the training set, so that it could theoretically be used as a factor for outcome prediction.

7 Evaluation Metrics

Spearman and Kendal Correlation Coefficients. In order to compare the importance scores attributed to the inputs and concepts by the different interpretability methods, we run each interpretability method on the inputs and evaluate the predictions with respect to the Silver Rationales annotations. We run a correlation study using Spearman (1904) and Kendall (1938) rank correlation coefficients and calculate statistical significance using a T-test. The differences assigned by the Contrastive

Annotation	Description	# of Cases Containing Annotation
Facts	A description of the case	11 000
Allegedly Violated Articles	A binary label indicating whether the lawyer claimed the article to be violated	11 000
Violated Articles	A binary label indicating whether the judge deemed the article violated	11 000
Positive/Negative/Null Cases	A three way label indicating whether the case was claimed and violated, not claimed, or claimed but not violated, respectively	11 000
Silver Allegation Rationales	Sentences from Facts referred to by the judge in the ruling	2 770
Gold Allegation Rationales	Sentences from Facts annotated by an expert as important	50
Legal Arguments	The presence of an argument	373
Legal Reasoning Concepts	The presence of a legal reasoning concept	735

Table 3: Data Statistics

Explanations method and the attributions of Integrated Gradients are treated as ranks.

Change in Accuracy and F1 scores. The method for evaluating the importance of legal concepts to the outcome prediction model is comparing the outcome prediction performance with and without the erasure of a given concept. We compare both accuracy and F1 scores of the predictions pre- and post-projection. In addition, we run a T-test to determine whether the predictions made pre- and post-projection are statistically significant.

8 Results

This section presents the quantitative and qualitative results of the interpretability methods. In order to ensure that the results cannot be accounted for by the short input sequence length of the LEGAL-BERT model (Chalkidis et al., 2020) or the ternary setup of the task (Valvoda et al., 2023), we replicate the results with the Longformer model (Beltagy et al., 2020) as well as the binary setup. The results of these experiments are presented in appendix C.

8.1 Integrated Gradients vs. Contrastive Explanations for Model Interpretability

A correlation method is applied to the results of the Integrated Gradients and Contrastive Explanations methods. In order to control for random effects, we compare to a random baseline, wherein the importance scores are assigned randomly to the inputs, within the same range as the scores of the interpretability methods.

The input paragraphs selected by both interpretability methods do not correlate with the Silver Rationales when looking at all the classes of the main task together, nor broken down by class. The effect size is very small, indicating that the models might not be relying on the information contained in the rationales for their predictions.

The breakdown between different facts and foils in Table 5 shows negligible correlations in all combinations of fact and foil.

Class	Method	Spearman		Kendall	
		coeff.	p-value	coeff.	p-value
null	IG	-0.036	**0.000**	-0.034	**0.000**
	CE	0.041	0.363	0.035	0.363
	random	0.011	0.490	0.009	0.491
pos	IG	-0.003	0.758	-0.003	0.758
	CE	0.088	**0.000**	0.070	**0.000**
	random	-0.016	0.362	-0.013	0.362
neg	IG	0.018	**0.000**	0.017	**0.000**
	CE	0.094	0.233	0.051	0.466
	random	0.045	0.511	0.042	0.511

Table 4: Results of the correlation study between the importance scores from Integrated Gradients (IG), Contrastive Explanations (CE) and random baseline on the one hand, and the Silver Rationales annotations on the other hand. The class refers to the ground truth. Statistically significant (p-value < 0.05) correlations are in bold.

To illustrate the types of paragraphs selected by the models as important, we look at one case in detail. Namely, in a case concerning the custody of a child, the lawyer has claimed that Articles 6 (Right to a fair trial) and 8 (Right to respect for private and family life) have been breached. However, the judge ruled that only Article 8 was violated, but not Article 6, meaning that the applicant is considered to have had a fair trial. Hence, this item has a negative label for Article 6, a positive label for Article 8, and null labels for all other articles. The model correctly predicts the negative label for Article 6.

The Contrastive Explanations method, using the positive label as a foil, lists the following sequence as the most important for this prediction: '*On 17 May 2005 the court dismissed the request for new access arrangements as the first applicant had failed to submit the required documents. It seems, however, that this decision did not become final as on 25 May 2005 the first applicant successfully requested that the proceedings be joined to pro-*

Fact	Foil	Method	Spearman coeff.	Spearman p-value	Kendall coeff.	Kendall p-value
null	pos	CE	0.020	**0.014**	0.017	**0.014**
		random	0.001	0.864	0.001	0.864
null	neg	CE	-0.017	**0.04**	-0.014	**0.04**
		random	0.005	0.570	0.004	0.570
pos	null	CE	-0.047	**0.000**	-0.039	**0.000**
		random	0.001	0.912	0.001	0.912
pos	neg	CE	-0.065	**0.000**	-0.053	**0.000**
		random	-0.007	0.428	-0.006	0.428
neg	null	CE	0.013	0.424	0.010	0.424
		random	-0.009	0.580	-0.007	0.580
neg	pos	CE	-0.062	**0.000**	-0.051	**0.000**
		random	-0.004	0.802	-0.003	0.802

Table 5: Results of the correlation study between the importance scores from Contrastive Explanations (CE) and the Silver Rationales annotations, compared to a random baseline, and broken down by fact and foil. Statistically significant (p-value < 0.05) correlations are in bold.

ceedings P 667/2003 (see paragraph 30 above)'.[6] This paragraph is also highlighted in the Silver Rationales data as one of the important factors for the case. This paragraph highlights that the applicant had failed to follow the required procedures for the trial, which is an argument as to why the court dismissing the request for new access arrangements is not deemed unlawful. This contrastive explanation indeed focuses on the reason why the claim was dismissed, rather than the reason why the claim was made in the first place. However, the paragraph selected as the second most important by the contrastive method is '*According to letters addressed to the court by the Šentjur Centre on 8 September 2003 and 3 May 2004, in the context of proceedings no. P 667/2003, the Šentjur Centre and the Unit attempted to organise supervised meetings between the applicants, but M.E. refused to cooperate.*'. As opposed to the first paragraph, these facts portray the reasons for accepting the claim, highlighting the refusal to cooperate of the second applicant, which could be interpreted as a breach of the right to a fair trial. These importance scores are contradictory to each other, both supporting and undermining the outcome.

Similarly, the Integrated Gradients method assigns the highest importance score to the following paragraph: '*On 1 August 2001 the Šentjur Centre issued an order granting the first applicant four hours a week with the second applicant, taking into account the expert committee's opinion and the fact that, at the supervised meeting between the applicants, the second applicant had not appeared to be afraid of the first applicant but, on the contrary, pleased to see him. The Šentjur Centre did not follow the first applicant's proposal that he should be allowed to pick the second applicant up at her nursery; instead it ordered M.E. to bring the second applicant to a meeting point at a local train station.*' This part of the input emphasizes the reasons for accepting the claim and assigning it a positive outcome, since the second applicant appears to be pleased to see the first applicant (the claimant). This could be interpreted as reasons to deem the trial unfair, as the text points to the circumstances in favour of the first applicant.

All in all, we observe through the qualitative analysis that the paragraphs selected by the interpretability methods appear to be relevant facts for the case, however not necessarily contributing to the predicted label. This suggests that they might not be particularly useful to an end user, given that they provide arguments for different outcomes to the predicted one.

8.2 Concept Erasure

The results of the Concept Erasure method on both the Legal Argument Mining (Habernal et al., 2023) and the Legal Reasoning Factors (Mumford et al., 2023) datasets are presented in Table 6. We confirm that the concepts are erased from the representation by observing that the Concept Prediction model performs at chance level, matching the majority accuracy, after the concept erasure. The F1 scores of the Concept Prediction task are often low even before the projection, however this matches the reportedly low legal concept prediction scores of Mumford et al. (2023) and Habernal et al. (2023). In order to ensure that concept erasure is happening, we perform the T-test on the predictions of the concept classification model before and after the projection. We find that in about half of the cases, the projection makes a significant difference to the predictions ($p<0.05$).

Overall, the results indicate that legal concepts are not absolutely necessary for the Outcome Prediction model, as the model performance is not significantly affected by the erasure of any of the legal concepts from both datasets. That is, the model with erased legal concepts is able to perform the task on par, or in some cases even better, than prior to the erasure.

[6] A larger subset of the facts from the case are presented in appendix A.

Concept Erased	Concept Task					Outcome Task			Concept Task					Outcome Task		
	Acc			F1		F1			Acc			F1		F1		
	maj	pre	post	pre	post	null	pos	neg	maj	pre	post	pre	post	null	pos	neg
None	–	–	–	–	–	.91	.55	.13	–	–	–	–	–	.93	.35	.39
	Legal Argument Mining (Habernal et al., 2023)															
	Development Set								**Test Set**							
1	.79	.79	.79	.25	.00	.93	.58	.40	.91	.86	.91	.40	.00	.92	.29	.13
2	.89	.86	.89	.33	.00	.93	.58	.50	.91	**.91**	**.91**	**.67**	**.00**	.92	.29	.19
3	.51	**.55**	**.48**	**.55**	**.29**	.93	.61	.46	.50	**.60**	**.50**	**.61**	**.15**	.92	.29	.13
4	.97	.97	.97	.00	.00	.93	.56	.50	.91	.91	.91	.00	.00	.92	.29	.13
5	.72	.79	.62	.84	.74	.93	.56	.47	.73	.63	.63	.73	.75	.92	.29	.13
6	.59	.52	.55	.50	.31	.93	.58	.50	.59	.59	.59	.53	.40	.92	.29	.13
7	.52	**.62**	**.41**	**.67**	**.56**	.93	.56	.46	.59	**.50**	**.36**	**.52**	**.53**	.92	.30	.20
	Legal Reasoning Factors (Mumford et al., 2023)															
	Development Set								**Test Set**							
1	.52	**.45**	**.51**	**.27**	**.00**	.91	.55	.13	.75	**.72**	**.75**	**.47**	**.00**	.92	.33	.39
2	.86	.83	.86	.00	.00	.91	.55	.12	.91	.91	.91	.00	.00	.93	.42	.40
3	.62	**.67**	**.62**	**.53**	**.00**	.92	.57	.13	.97	1.00	.97	1.00	.00	.92	.35	.38
4	.90	**.67**	**.90**	**.00**	**.00**	.92	.55	.06	.84	**.84**	**.84**	**.55**	**.00**	.93	.19	.39
5	.83	.83	.83	.00	.00	.91	.52	.06	.88	.84	.88	.00	.00	.93	.27	.38
6	.97	.97	.97	.00	.00	.92	.55	.12	.97	.97	.97	.00	.00	.93	.35	.39
7	.79	**.62**	**.79**	**.00**	**.00**	.92	.55	.13	.75	**.75**	**.75**	**.33**	**.00**	.93	.35	39
8	.83	**.79**	**.83**	**.88**	**.91**	.91	.48	.19	.75	**.81**	**.75**	**.88**	**.86**	.92	.36	.41
9	.62	**.83**	**.62**	**.78**	**.00**	.92	.55	.07	.75	**.72**	**.75**	**.40**	**.00**	.92	.29	.43
10	.83	**.86**	**.83**	**.60**	**.00**	.92	.48	.13	.78	.78	.78	.36	.00	.93	.27	.38
11	.93	.93	.93	.00	.00	.92	.55	.13	.94	.94	.94	.00	.00	.93	.35	.43
12	.76	**.86**	**.76**	**.67**	**.00**	.92	.55	.13	.88	**.72**	**.88**	**.00**	**.00**	.93	.35	.39
13	.90	**.76**	**.90**	**.00**	**.00**	.92	.55	.13	.84	**.66**	**.84**	**.00**	**.00**	.93	.38	.42
14	.90	.90	.86	.00	.00	.91	.55	.13	.88	.84	.88	.00	.00	.93	.35	.39
15	.90	.90	.90	.00	.00	.92	.55	.13	.97	.94	.97	.00	.00	.93	.36	.39
16	.93	**.86**	**.93**	**.33**	**.00**	.92	.52	.13	.75	**.72**	**.75**	**.30**	**.00**	.93	.34	.39
17	.97	.93	.97	.00	.00	.92	.55	.13	.94	.91	.94	.00	.00	.93	.35	.39
18	.76	.72	.76	.00	.00	.91	.55	.13	.84	.81	.84	.00	.00	.93	.35	.39
19	.97	.97	.97	.00	.00	.91	.55	.13	.97	.97	.97	.00	.00	.93	.35	.39
20	.79	**.69**	**.79**	**.31**	**.00**	.92	.52	.13	.66	**.75**	**.66**	**.63**	**.00**	.92	.36	.37

Table 6: Results of the Concept Erasure method: accuracy and F1 scores of the outcome prediction model as well as concept prediction model pre- and post-projection, including a majority class baseline for concept prediction. Statistically significant differences between pre- and post-projection predictions are marked in bold. The concepts are listed in appendix B.

9 Conclusion

We have studied three interpretability methods in the domain of legal outcome prediction. Our experimental results show a small variance in the correlation between the importance scores assigned by different interpretability methods and the ground truth. Worryingly, even removing the information a lawyer would consider essential for reasoning over the data has an insiginificant effect on the model performance. We interpret this result as a call for caution in using automated legal outcome prediction models as they do not appear to be grounded in legal reasoning to the extent that would be necessary for ensuring reliability.

Future work in NLLP should continue searching for ways to make legal outcome prediction models more transparent by investigating their legal expertise. Studying why predicting negative outcome prediction remains a difficult task is only one direction of such research. New directions could involve the study of biases that may be affecting the decision making of the models.

Limitations

This study is limited to only English language data. In future work, it should be extended to other languages as well as other jurisdictions. As far as the results of the study are concerned, the outputs of the explainability method depend on the performance of the outcome prediction model, which could itself be improved, especially on the negative case in the ternary setup. While we acknowledge that the explainability suffers from model errors, this is in line with the argument that improvements to the model should incorporate interpretable legal reasoning.

Ethics Statement

Our research indicates it is not safe to deploy outcome prediction models to the real world, as the predictions of the model do not have a strong basis in legal reasoning and therefore may be biased through dependence on spurious correlations.

Acknowledgments

Ieva Raminta Staliūnaitė is supported by Huawei. Josef Valvoda is funded by the Nordic Programme for Interdisciplinary Research Grant 105178 and the Danish National Research Foundation Grant no. DNRF169.

References

Vincent AWMM Aleven. 1997. *Teaching case-based argumentation through a model and examples*. Ph.D. Dissertation, University of Pittsburgh.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Kevin D Ashley. 1991. Reasoning with cases and hypotheticals in hypo. *International journal of man-machine studies*, 34(6):753–796.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv e-prints*, pages arXiv–2004.

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*, 29:213–238.

Karl Branting, Brandy Weiss, Bradford Brown, Craig Pfeifer, A Chakraborty, Lisa Ferro, Mark Pfaff, and Alex Yeh. 2019. Semi-supervised methods for explainable legal prediction. In *Proceedings of the seventeenth international conference on artificial intelligence and law*, pages 22–31.

Stefanie Brüninghaus and Kevin D Ashley. 2006. Progress in textual case-based reasoning: predicting the outcome of legal cases from text. In *AAAI*, pages 1577–1580.

Oana-Maria Camburu. 2020. *Explaining Deep Neural Networks*. Ph.D. thesis, University of Oxford.

Lang Cao, Zifeng Wang, Cao Xiao, and Jimeng Sun. 2024. Pilot: Legal case outcome prediction with case law. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 609–621.

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.

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 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Mexico City, Mexico. Association for Computational Linguistics.

Joe Collenette, Katie Atkinson, and Trevor Bench-Capon. 2020. An explainable approach to deducing outcomes in european court of human rights cases using adfs. In *Computational Models of Argument*, pages 21–32. IOS Press.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*.

Qian Dong and Shuzi Niu. 2021. Legal judgment prediction via relational learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 983–992.

Jakub Drápal, H Westermann, J Savelka, et al. 2023. Using large language models to support thematic analysis in empirical legal studies. *Legal Knowledge and Information Systems*, pages 197–206.

Mohammad H Falakmasir and Kevin D Ashley. 2017. Utilizing vector space models for identifying legal factors from text. In *Legal Knowledge and Information Systems*, pages 183–192. IOS Press.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664.

Morgan Gray, Jaromir Savelka, Wesley Oliver, and Kevin Ashley. 2023. Automatic identification and empirical analysis of legally relevant factors. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 101–110.

Morgan A Gray, Jaromir Savelka, Wesley M Oliver, and Kevin D Ashley. 2024. Empirical legal analysis simplified: reducing complexity through automatic identification and evaluation of legally relevant factors. *Philosophical Transactions of the Royal Society A*, 382(2270):20230155.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2023. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, pages 1–38.

Philipp Hacker, Ralf Krestel, Stefan Grundmann, and Felix Naumann. 2020. Explainable ai under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law*, 28:415–439.

Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, et al. 2020. Robustness and explainability of artificial intelligence. *Publications Office of the European Union*, 207:2020.

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.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of NAACL-HLT*, pages 3543–3556.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *Preprint*, arXiv:2009.07896.

Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Shangbang Long, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Automatic judgment prediction via legal reading comprehension. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 558–572. Springer.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

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.

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

Emre Mumcuoğlu, Ceyhun E Öztürk, Haldun M Ozaktas, and Aykut Koç. 2021. Natural language processing in law: Prediction of outcomes in the higher courts of turkey. *Information Processing & Management*, 58(5):102684.

Jack Mumford, Katie Atkinson, and Trevor Bench-Capon. 2023. Combining a legal knowledge model with machine learning for reasoning with legal cases. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 167–176.

Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, and Sally Gao. 2021. Towards explainable ai: Assessing the usefulness and impact of added explainability features in legal document summarization. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. 2021. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022.

Vihari Piratla, Juyeon Heo, Sukriti Singh, and Adrian Weller. 2023. Estimation of concept explanations should be uncertainty aware. *arXiv preprint arXiv:2312.08063*.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. 2022. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR.

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.

Olivier Salaün, Fabrizio Gotti, Philippe Langlais, and Karim Benyekhlef. 2022. Why do tenants sue their landlords? answers from a topic model. In *Legal Knowledge and Information Systems*, pages 113–122. IOS Press.

Tyss Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022. Deconfounding legal judgment prediction for european court of human rights cases towards better alignment with experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1138.

Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242.

Gil Semo, DorBernsohn BenHagag GilaHayat, and Joel Niklaus. 2022. Classactionprediction: A challenging benchmark for legal judgment prediction of class action cases in the us. *NLLP 2022*, 2022:31–46.

Rafe Athar Shaikh, Tirath Prasad Sahu, and Veena Anand. 2020. Predicting outcomes of legal cases based on legal factors using classifiers. *Procedia Computer Science*, 167:2393–2402.

Jerrold Soh Tsin Howe. 2024. Discovering significant topics from legal decisions with selective inference. *Philosophical Transactions of the Royal Society A*, 382(2270):20230147.

C Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal judgement prediction for uk courts. In *Proceedings of the 3rd International Conference on Information Science and Systems*, pages 204–209.

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

Josef Valvoda and Ryan Cotterell. 2024. Towards explainability in legal outcome prediction models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7262–7282.

Josef Valvoda, Ryan Cotterell, and Simone Teufel. 2023. On the Role of Negative Precedent in Legal Outcome Prediction. *Transactions of the Association for Computational Linguistics*, 11:34–48.

Josef Valvoda, Alec Thompson, Ryan Cotterell, and Simone Teufel. 2024. The ethics of automating legal actors. *Transactions of the Association for Computational Linguistics*, 12:700–720.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Michael Benedict L Virtucio, Jeffrey A Aborot, John Kevin C Abonita, Roxanne S Avinante, Rother Jay B Copino, Michelle P Neverida, Vanesa O Osiana, Elmer C Peramo, Joanna G Syjuco, and Glenn Brian A Tan. 2018. Predicting decisions of the philippine supreme court using natural language processing and machine learning. In *2018 IEEE 42nd annual computer software and applications conference (COMPSAC)*, volume 2, pages 130–135. IEEE.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. Gradient-based analysis of nlp models is manipulable. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 247–258.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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)*. Association for Computational Linguistics.

Shanshan Xu, Santosh T.y.s.s, Oana Ichim, Isabella Risini, Barbara Plank, and Matthias Grabmair. 2023. From dissonance to insights: Dissecting disagreements in rationale construction for case outcome classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9576, Singapore. Association for Computational Linguistics.

Hiroaki Yamada, Takenobu Tokunaga, Ryutaro Ohara, Akira Tokutsu, Keisuke Takeshita, and Mihoko Sumida. 2024. Japanese tort-case dataset for rationale-supported legal judgment prediction. *Artificial Intelligence and Law*, pages 1–25.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565.

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.

A The Facts of Case 001-118248

‘4. The first applicant, Mr Eberhard, was born in 1968 and lives in Ponikva. The second applicant, M., is his daughter.’, ‘5. On 8 April 2001 the first applicant’s wife, M.E., together with the second applicant, then aged four, moved out of the flat in which they had been living with the first applicant. M.E. subsequently filed a petition for divorce.’, ‘6. On 4 May 2001 the first applicant and his wife, with whom the second applicant was living, signed an agreement on access arrangements.’, ‘7. On 12 June 2001 the first applicant filed a request with the Šentjur Social Welfare Centre (“the Šentjur Centre”) seeking formal determination of the access arrangements, claiming that since 12 May 2001 M.E. had denied him access to the second applicant.’, ‘8. During the following month M.E. gave a number of statements at the Šentjur Centre, opposing contact between the applicants, stating that the first applicant represented a danger to her and the second applicant. She also lodged a criminal complaint against the first applicant for endangering their safety.’, [...] ‘11. On 1 August 2001 the Šentjur Centre issued an order granting the first applicant four hours a week with the second applicant, taking into account the expert committee’s opinion and the fact that, at the supervised meeting between the applicants, the second applicant had not appeared to be afraid of the first applicant but, on the contrary, pleased to see him. The Šentjur Centre did not follow the first applicant’s proposal that he should be allowed to pick the second applicant up at her nursery; instead it ordered M.E. to bring the second applicant to a meeting point at a local train station.’, [...] ‘19. On 15 June 2004 the Ministry quashed the impugned enforcement orders, finding that M.E. had not been informed of the first applicant’s notices concerning non-compliance and had had no opportunity of participating in the proceedings and presenting arguments in her favour. [...] ‘24. On 6 June 2003 the first applicant lodged an application for custody of the second applicant, relying on the fact that M.E. was denying them contact. He also requested an interim order under which the second applicant would be placed in his custody pending the outcome of the proceedings, and the appointment of a curator ad litem to represent the second applicant’s interests in the proceedings. [...] ‘29. However, as M.E. continued to refuse any contact between the applicants, on 16 August 2004 the first applicant requested that the proceedings be resumed and a hearing was scheduled for 7 October 2004. It was adjourned as the court decided, further to the first applicant’s request, to appoint an expert psychologist. On 19 October 2004 the court appointed expert D.T. to produce an opinion in the case.’, [...] In addition, the first applicant alerted the court to the fact that he had had no access to the second applicant in the past four and a half years, except on one occasion at her school.’, ‘32. In the meantime, the appointed expert informed the court on 22 September 2005 that he was unable to prepare the opinion as M.E. had refused to cooperate. [...] Subsequently, on 26 May 2006, the court issued a decision rejecting the first applicant’s application for provisional custody and upholding his alternative request for an interim access order. [...] ‘45. On 2 March 2007 the first applicant lodged a supervisory appeal, relying on section 6 of the Act on Protection of the Right to a Hearing without Undue Delay (“the 2006 Act”).’, [...] ‘62. On 17 May 2005 the court dismissed the

request for new access arrangements as the first applicant had failed to submit the required documents. It seems, however, that this decision did not become final as on 25 May 2005 the first applicant successfully requested that the proceedings be joined to proceedings P 667/2003 (see paragraph 30 above).'

B Legal Concepts Used in this Study

The concepts from the Mumford et al. (2023) dataset:

1. Access to Court
2. Allowed Time and Facilities for Defence
3. Allowed to Defend in Person or Through Legal Assistance
4. Allowed to Fairly Examine Witnesses
5. Balance of Complexity and Circumstance
6. Conducted Publicly Where Appropriate
7. Equality of Arms and Adversarial Hearing
8. Fair
9. Had the Minimum Rights
10. Independent and Impartial
11. Informed Promptly
12. Integrity of Evidence
13. Legal Certainty is Upheld
14. No Adverse Effect from Alternative Proceedings
15. No adverse Prejudicial Statements
16. No Unreasonable Delays
17. Option of Free Access to Interpreter
18. Presumption of Innocence
19. Public Hearing
20. Reasonable Time

The concepts from the Habernal et al. (2023) dataset:

1. Distinguishing
2. Scope of Assessment
3. Consensus of the Procedural Parties
4. Meaning & Purpose Interpretation
5. Proportionality Test - Appropriateness
6. Proportionality Test - Legitimate Purpose
7. Proportionality Test - Legal Basis

C Longformer and Binary Setup Results

C.1 Longformer Three Way Classification Model

Tables 7 and 8 present the results of the Contrastive Explanation, Integrated Gradients and Concept Erasure on the Longformer three way outcome prediction model. The results corroborate the results seen in Section 8, namely low correlation scores between contrastive explanation and integrated gradients importance scores against silver rationales, and unchanged outcome prediction scores after concept erasure. Given the lack of difference between the results with the Longformer model and the LEGAL-BERT model, we conclude that the effect observed in this study is not due to some idiosyncratic behavior of LEGAL-BERT.

Class	Method	Spearman		Kendall	
		coeff.	p-value	coeff.	p-value
null	IG	-0.028	**0.000**	-0.026	**0.000**
	CE	0.011	0.180	0.009	0.180
	random	0.003	0.733	0.002	0.733
pos	IG	0.010	0.303	0.009	0.303
	CE	-0.060	**0.000**	-0.049	**0.000**
	random	-0.009	0.322	-0.007	0.322
neg	IG	-0.019	0.205	-0.018	0.205
	CE	-0.039	**0.013**	-0.032	**0.013**
	random	0.001	0.926	0.001	0.926

Table 7: Results of the correlation study between the importance scores from Integrated Gradients (IG), Contrastive Explanations (CE) and random baseline on the one hand, and the Silver Rationales annotations on the other hand. The class refers to the ground truth. Statistically significant (p-value < 0.05) correlations are in bold.

C.2 LEGAL-BERT Binary Classification Model

Tables 9 and 10 present the results of the Contrastive Explanation, Integrated Gradients and Concept Erasure on the LEGAL-BERT binary outcome prediction model. The results corroborate the results seen in Section 8, namely low correlation

Erased Concept	Concept Task					Outcome Task			Concept Task					Outcome Task		
	Acc			F1		F1			Acc			F1		F1		
	maj	pre	post	pre	post	null	pos	neg	maj	pre	post	pre	post	null	pos	neg
None	–	–	–	–	–	.93	.57	.36	–	–	–	–	–	.93	.38	.15
	Legal Argument Mining (Habernal et al., 2023)															
	Development Set								**Test Set**							
1	.79	.86	.76	.60	.22	.93	.57	.36	.91	.68	.86	.00	.40	.94	.38	.15
2	.90	.86	.90	.33	.00	.93	.57	.39	.91	**.82**	**.91**	**.33**	**.00**	.94	.38	.16
3	.52	**.48**	**.52**	**.35**	**.00**	.93	.57	.36	.50	**.50**	**.50**	**.42**	**.00**	.94	.38	.21
4	..97	.97	.97	.00	.00	.93	.57	.39	.91	.91	.91	.00	.00	.95	.38	.17
5	.72	**.72**	**.72**	**.79**	**.84**	.94	.57	.41	.73	**.68**	**.73**	**.76**	**.84**	.94	.38	.21
6	.59	**.59**	**.59**	**.46**	**.00**	.93	.55	.37	.59	**.64**	**.59**	**.43**	**.00**	.94	.38	.22
7	.52	.66	.59	.67	.63	.94	.57	.41	.59	.41	.27	.48	.43	.94	.38	.16
	Legal Reasoning Factors (Mumford et al., 2023)															
	Development Set								**Test Set**							
1	.52	**.52**	**.52**	**.30**	**.00**	.93	.45	.27	.75	**.66**	**.75**	**.15**	**.00**	.94	.29	.50
2	.86	.86	.83	.33	.00	.93	.48	.26	.91	.88	.91	.00	.00	.94	.27	.46
3	.62	**.69**	**.62**	**.47**	**.00**	.93	.48	.26	.97	.94	.97	.00	.00	.94	.29	.50
4	.90	**.79**	**.90**	**.40**	**.00**	.92	.48	.25	.84	**.78**	**.84**	**.22**	**.00**	.94	.32	.50
5	.83	**.76**	**.83**	**.22**	**.00**	.93	.50	.25	.88	**.78**	**.88**	**.36**	**.00**	.94	.29	.50
6	.97	.93	.97	.50	.00	.93	.48	.26	.97	.88	.94	.00	.00	.94	.30	.50
7	.79	**.66**	**.79**	**.17**	**.00**	.92	.50	.24	.75	**.56**	**.75**	**.13**	**.00**	.94	.32	.54
8	.83	**.66**	**.83**	**.78**	**.91**	.93	.48	.26	.75	**.59**	**.75**	**.70**	**.86**	.94	.29	.50
9	.62	.76	.79	.67	.75	.93	.48	.26	.75	.75	.75	.33	.56	.93	.32	.51
10	.83	**.66**	**.79**	**.17**	**.00**	.93	.50	.25	.78	.69	.78	.17	.22	.94	.32	.45
11	.93	.97	.93	.67	.00	.93	.48	.26	.94	.88	.94	.00	.00	.94	.29	.42
12	.76	**.90**	**.76**	**.80**	**.00**	.92	.48	.25	.88	**.81**	**.88**	**.25**	**.00**	.94	.33	.50
13	.90	**.79**	**.90**	**.00**	**.00**	.93	.50	.26	.84	**.75**	**.84**	**.20**	**.00**	.93	.32	.46
14	.90	.90	.90	.40	.00	.92	.48	.25	.88	**.72**	**.88**	**.00**	**.00**	.94	.30	.46
15	.90	.93	.90	.50	.00	.92	.48	.25	.97	1.00	.97	.67	.00	.94	.32	.45
16	.93	**.66**	**.93**	**.17**	**.00**	.92	.50	.24	.75	**.66**	**.75**	**.35**	**.00**	.94	.29	.47
17	.97	.97	.97	.67	.00	.93	.48	.26	.94	.94	.94	.00	.00	.94	.30	.46
18	.76	**.79**	**.76**	**.50**	**.00**	.93	.48	.26	..84	.84	.84	.29	.00	.94	.30	.42
19	.97	.90	.97	.40	.00	.93	.48	.26	.97	.94	.97	.50	.00	.94	.29	.46
20	.79	.69	.79	.47	.57	.93	.48	.26	.66	.72	.50	.61	.33	.94	.32	.54

Table 8: Results of the Concept Erasure method with the Longformer three way outcome prediction model: accuracy and F1 scores of the outcome prediction model as well as concept prediction model pre- and post-projection, including a majority class baseline for concept prediction. Statistically significant differences between pre- and post-projection predictions are marked in bold. The concepts are listed in appendix B.

between the importance scores assigned by contrastive explanation and integrated gradients methods against silver rationales, as well as no change in outcome prediction performance after concept erasure. Based on the lack of difference between the results observed with the ternary and binary setups, we conclude that the results cannot be accounted for by the more difficult three way classification task.

Class	Method	Spearman		Kendall	
		coeff.	p-value	coeff.	p-value
neg	IG	0.041	**0.000**	0.038	**0.000**
	CE	0.011	0.483	0.083	0.483
	random	0.010	0.371	0.008	0.371
pos	IG	0.015	0.388	0.014	0.388
	CE	0.026	0.367	0.021	0.367
	random	0.028	0.411	0.073	0.411

Table 9: Results of the correlation study between the importance scores from Integrated Gradients (IG), Contrastive Explanations (CE) and random baseline on the one hand, and the Silver Rationales annotations on the other hand. The class refers to the ground truth. Statistically significant (p-value < 0.05) correlations are in bold.

Concept Erased	Concept Task					Outcome Task	Concept Task					Outcome Task
		Acc		F1		F1		Acc		F1		F1
	maj	pre	post	pre	post		maj	pre	post	pre	post	
None	–	–	–	–	–	.76	–	–	–	–	–	.76
	Legal Argument Mining (Habernal et al., 2023)											
	Development Set						**Test Set**					
1	.79	.79	.79	.25	.00	.76	.93	.86	.91	.40	.00	.75
2	.90	.86	.90	.33	.00	.76	.93	**.91**	**.91**	**.00**	**.00**	.75
3	.52	**.55**	**.48**	**.55**	**.29**	.78	.50	**.59**	**.50**	**.61**	**.15**	.75
4	.97	.97	.97	.00	.00	.78	.91	.91	.91	.00	.00	.75
5	.72	.79	.62	.84	.74	.78	.73	.64	.64	.73	.75	.75
6	.59	.52	.55	.50	.38	.78	.59	.59	.59	.53	.40	.75
7	.52	**.62**	**.41**	**.67**	**.56**	.78	.59	**.50**	**.36**	**.52**	**.53**	.75
	Legal Reasoning Factors (Mumford et al., 2023)											
	Development Set						**Test Set**					
1	.52	**.45**	**.52**	**.27**	**.00**	.76	.75	**.72**	**.75**	**.47**	**.00**	.75
2	.86	.83	.86	.00	.00	.76	.91	.91	.91	.00	.00	.76
3	.62	.69	.62	.53	.00	.76	.97	**1.00**	**.97**	**1.00**	**.00**	.76
4	.90	**.69**	**.90**	**.00**	**.00**	.77	.84	**.84**	**.84**	**.55**	**.00**	.76
5	.83	.83	.83	.00	.00	.76	.88	.84	.88	.00	.00	.76
6	.97	.97	.97	.00	.00	.77	.97	.97	.97	.00	.00	.76
7	.79	**.62**	**.79**	**.00**	**.00**	.77	.75	**.75**	**.75**	**.33**	**.00**	.76
8	.83	**.79**	**.83**	**.88**	**.91**	.76	.75	**.81**	**.75**	**.88**	**.86**	.75
9	.62	**.83**	**.62**	**.78**	**.00**	.77	.75	**.72**	**.75**	**.40**	**.00**	.75
10	.83	**.86**	**.83**	**.60**	**.00**	.77	.78	**.78**	**.78**	**.36**	**.00**	.76
11	.93	.93	.93	.00	.00	.77	.94	.93	.93	.00	.00	.76
12	.76	**.86**	**.76**	**.67**	**.00**	.77	.88	**.72**	**.88**	**.00**	**.00**	.76
13	.90	**.76**	**.90**	**.00**	**.00**	.77	.84	**.66**	**.84**	**.00**	**.00**	.76
14	.90	.90	.86	.00	.00	.76	.88	.84	.88	.00	.00	.76
15	.90	.90	.90	.00	.00	.77	.97	.94	.97	.00	.00	.76
16	.93	**.86**	**.93**	**.33**	**.00**	.77	.75	**.72**	**.75**	**.31**	**.00**	.76
17	.97	.93	.97	.00	.00	.77	.94	.91	.94	.00	.00	.76
18	.76	.72	.76	.00	.00	.76	.84	.81	.84	.00	.00	.76
19	.97	.97	.97	.00	.00	.76	.97	.97	.97	.00	.00	.75
20	.79	**.69**	**.79**	**.31**	**.00**	.77	.66	**.75**	**.66**	**.63**	**.00**	.75

Table 10: Results of the Concept Erasure method with the LEGAL-BERT binary outcome prediction model: accuracy and F1 scores of the outcome prediction model as well as concept prediction model pre- and post-projection, including a majority class baseline for concept prediction. Statistically significant differences between pre- and post-projection predictions are marked in bold. The concepts are listed in appendix B.