# The Craft of Selective Prediction: Towards Reliable Case Outcome Classification - An Empirical Study on European Court of Human Rights Cases

**Santosh T.Y.S.S, Irtiza Chowdhury, Shanshan Xu, Matthias Grabmair**
School of Computation, Information, and Technology;
Technical University of Munich, Germany

## Abstract

In high-stakes decision-making tasks within legal NLP, such as Case Outcome Classification (COC), quantifying a model's predictive confidence is crucial. Confidence estimation enables humans to make more informed decisions, particularly when the model's certainty is low, or where the consequences of a mistake are significant. However, most existing COC works prioritize high task performance over model reliability. This paper conducts an empirical investigation into how various design choices—including pre-training corpus, confidence estimator and fine-tuning loss—affect the reliability of COC models within the framework of selective prediction. Our experiments on the multi-label COC task, focusing on European Court of Human Rights (ECtHR) cases, highlight the importance of a diverse yet domain-specific pre-training corpus for better calibration. Additionally, we demonstrate that larger models tend to exhibit overconfidence, Monte Carlo dropout methods produce reliable confidence estimates, and confident error regularization effectively mitigates overconfidence. To our knowledge, this is the first systematic exploration of selective prediction in legal NLP. Our findings underscore the need for further research on enhancing confidence measurement and improving the trustworthiness of models in the legal domain.

## 1 Introduction

The task of Case Outcome Classification (COC) involves categorizing the outcome of a legal case based on the text of case facts. It has garnered substantial attention not only within the legal but also in the NLP community (Aletras et al., 2016; Chalkidis et al., 2019; Hwang et al., 2022). It is important to acknowledge that these tasks commonly labeled as 'Legal Judgment Prediction' are, in reality, instances of retrospective classification rather than prediction as the fact statements obtained from judgment documents are often not finalized until the decision outcome is known, as emphasized by Medvedeva et al. 2021, introducing the potential confounding artifacts in data (Santosh et al., 2022a). The main utility of this task and data lies in understanding the capabilities of NLP models to analyze fact statements for extracting and learning text patterns corresponding to specific convention articles, as drafted by the court. Identifying potentially violated human rights provisions from a textual fact description is a task that human experts can do well and that requires substantial domain knowledge along with textual understanding phenomenon (Chalkidis et al., 2022a). Correlation between text patterns and violations has been a subject of interest in empirical research, particularly in political science (Segal, 1984; Kort, 1957; Nagel, 1963).

Though COC has witnessed improvements in performance with use of better pre-trained models (Douka et al., 2021; Chalkidis et al., 2020; Xiao et al., 2021) or innovative modelling strategies with better architectures and loss functions (Tyss et al., 2023b; Yue et al., 2021; Zhao et al., 2022; Zhong et al., 2018) or incorporation of legal knowledge (Tyss et al., 2023a; Gan et al., 2021), the accuracy of these models is not guaranteed for all instances. Therefore, it is crucial to assess the reliability of model predictions, particularly in high-stakes decision-making scenarios—an aspect that has not received adequate attention in the community.

In this paper, we explore the reliability of COC systems using selective prediction setting (El-Yaniv et al., 2010). In this setting, the objective is to reduce the error rate by abstaining from predictions when the model is uncertain, while maintaining high coverage. In essence, we consider a model reliable if it possesses the self-awareness capability to acknowledge when it doesn't know enabling it to defer to humans for manual inspection, thereby ensuring its trustworthiness (Geifman and El-Yaniv, 2017). Under the selective prediction setting, we construct a selective classifier by



*Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3656–3674

combining a standard classifier with a confidence estimator. The confidence estimator gauges the model's confidence for a given input instance and based on this confidence, the selective classifier decides whether to abstain from predicting on uncertain cases. An ideal confidence estimator should provide higher confidence for correctly classified examples compared to incorrect ones.

In this study, we conduct an empirical investigation on the COC task, focusing on European Court of Human Rights (ECtHR) cases which adjudicates complaints by individuals against states regarding alleged violations of their rights as enshrined in the European Convention of Human Rights. Our goal is to assess four design choices to obtain more reliable COC models: (i) How does the choice of pre-trained models, such as general BERT (Kenton and Toutanova, 2019) or legal domain-specific models like LegalBERT (Chalkidis et al., 2019) or LexLM (Chalkidis et al., 2023), impact reliability? Does the size of the pre-trained model, such as Base or Large, play a role? (ii) Is there a universally effective confidence estimator, such as Softmax Response (Hendrycks and Gimpel, 2016) or Monte Carlo dropout based methods(Gal and Ghahramani, 2016)? (iii) How do additional training loss constraints in the form of regularizers (Xin et al., 2021; Shamsi et al.) or learning directly with abstention as an option (Liu et al., 2019) affect it? We evaluate these design choices on three COC task variants with varying difficulty, ranging from predicting violations alleged by the claimant to violations decided by the court. Additionally, we examine the impact of these choices on different buckets of articles based on frequency of cases with corresponding article violations.

Based on our empirical exploration, we observe: (i) Domain-specific pre-training enhances model calibration, but exclusive focus on downstream task-specific pre-training corpus is detrimental. A domain-related yet diverse corpus is crucial for effective pre-training. Larger models, while more accurate, tend to be overconfident. (ii) Computationally expensive Monte Carlo Dropout methods provide superior confidence estimates. (iii) Adding confident error regularization (Xin et al., 2021) improves model calibration. To encourage future work towards better uncertainty quantification in COC task, we release our code, including pipelines to evaluate design choices based on selective prediction and classification performance.

## 2 Related Work

**Selective Prediction** Selective prediction, in which a model can either predict or abstain on each test example, is a long-standing research area in machine learning (Chow, 1957; Hellman, 1970; Fumera and Roli, 2002; Cortes et al., 2016; El-Yaniv et al., 2010; Geifman and El-Yaniv, 2017). Selective prediction has recently received considerable attention from the NLP community on various tasks such as Quesrion answering (Kamath et al., 2020; Garg and Moschitti, 2021), classification and NLI (Gu and Hopkins, 2023; Varshney et al., 2022b,a), knowledge probing (Yoshikawa and Okazaki, 2023) and generation (Ren et al., 2022; Chen et al., 2023; Cole et al., 2023) and is mostly related to uncertainty/confidence estimation (Vazhentsev et al., 2022, 2023). Another related area to selective prediction, albeit remotely, is calibration (Jiang et al., 2018; Desai and Durrett, 2020; Wang et al., 2020; Guo et al., 2017) which deals with the development of interpretable confidence measures focusing on adjusting the overall confidence level of a model, while selective prediction is based on relative confidence among the examples. In this work, we employ selective prediction framework to evaluate the reliability of models in the context of multi-label COC task, in contrast to prior works that predominantly focused on single-label classification tasks.

**COC** COC has been explored using corpora from different jurisdictions, such as the ECtHR (Chalkidis et al., 2022a; Aletras et al., 2016; Medvedeva et al., 2021; Santosh et al., 2023a,b) Chinese Criminal Courts (Yue et al., 2021), US Supreme Court (Katz et al., 2017; Kaufman et al., 2019), Indian Supreme Court (Malik et al., 2021; Shaikh et al., 2020) French court of Cassation (Şulea et al., 2017b), Federal Supreme Court of Switzerland (Niklaus et al., 2021), UK courts (Strickson and De La Iglesia, 2020), German courts (Waltl et al., 2017), Brazilian courts (Lage-Freitas et al., 2022), the Philippine Supreme court (Virtucio et al., 2018), and the Thailand Supreme Court (Kowsrihawat et al., 2018). While early works relied on rule-based approaches (Segal, 1984; Nagel, 1963), later works used classification techniques using bag-of-words features (Aletras et al., 2016; Şulea et al., 2017a). Most recent work in COC use deep learning (Zhong et al., 2018, 2020; Yang et al., 2019) followed by adoption of pre-trained transformer models (Chalkidis et al., 2019; Niklaus et al., 2021), including legal-domain specific pre-

trained variants (Zheng et al., 2021; Chalkidis et al., 2023). Furthermore, different strategies were proposed by leveraging dependency between auxiliary tasks (Tyss et al., 2023b; Yue et al., 2021; Valvoda et al., 2023) or with additional loss such as contrastive learning (Tyss et al., 2023b; Zhang et al., 2023) or by injecting legal knowledge (Liu et al., 2023; Tyss et al., 2023a).

While the majority of existing research in COC focuses on improving predictive performance, there is an increasing emphasis on the reliability of models within legal NLP. This includes perspectives on explainability (Chalkidis et al., 2021; Santosh et al., 2022a; Xu et al., 2023) and fairness (Wang et al., 2021; Chalkidis et al., 2022b; Li et al., 2022; Baumgartner et al., 2024). Recently, mainstream NLP has seen researchers, such as Baan et al. 2024, propose that the overall reliability of a model is determined by two facets: 1) fairness, which is related to the alignment of model confidence with human expectations, and 2) trustworthiness, which involves accurate confidence measurement by the model. Xu et al. 2024 study models' calibration in alignment with human behavior using split-vote cases from ECtHR, pioneering on research on model reliability from a fairness perspective in legal NLP. In contrast, our work systematically explores selective prediction in COC task. To the best of our knowledge, this is the first COC research on model reliability from a trustworthiness perspective in Legal NLP.

## 3 ECtHR Tasks & Datasets

We chose to work with the ECtHR corpus because of its publicly available dataset with detailed article-specific allegations and violation information, leading to multi-label classification setting, in contrast to the simplified binary classification setting in corpora from other jurisdictions (Niklaus et al., 2021; Alali et al., 2021). We experiment with the following three COC task variants. Following Valvoda et al. 2023, we use the 14 articles which form the core rights of the convention.

**Task B: Allegation Identification** (Chalkidis et al., 2021) We utilize data from LexGLUE ECtHR B (Chalkidis et al., 2022a), where the fact description serves as input to identify the set of convention articles that the claimant alleges to have been violated.

**Task A: Violation Identification** (Chalkidis et al., 2019) We leverage data from LexGLUE ECtHR A to predict which of the convention's articles has been deemed violated by the court using the facts description as input. Task A is more challenging than B as it involves the identification of suitable articles along with prediction of their violations.

**Task A|B: Violation Identification given Allegation information** (Santosh et al., 2022b) This involves the identification of violations from the case facts along with allegedly violated articles as the input. This task mirrors the realistic legal process, as the court is aware of the allegations made by the applicants when determining the violations. This task is easier compared to Task A, as the first substep of Task A (i.e., identifying suitable articles) is provided directly as input in this variant.

**Dataset splits & Metrics for Prediction Performance** LexGLUE consists of of 11k case fact descriptions chronologically split into training (2001–2016, 9k cases), validation (2016–2017, 1k cases), and test sets (2017-2019, 1k cases). Following Chalkidis et al. 2022a, we report macro-F1 (m-F1) scores for all tasks across the 14 articles.

## 4 Selective Prediction

A standard classifier learns a function $f : X \rightarrow Y$, takes input $X$ and maps it to set of labels $Y$. We pair a standard classifier with a selection function $g : X \rightarrow \{0, 1\}$ to obtain a selective classifier $h = (f, g)$; $h : X \rightarrow Y \cup \{\perp\}$, $\perp$ is a special label indicating the abstention of prediction. Given an input $x$, the selective classifier outputs as follows:

$$h(x) = (f, g)(x) = \begin{cases} f(x) & \text{if } g(x) = 1 \\ \perp & \text{if } g(x) = 0 \end{cases}$$

The selective classifier yields an output from $f$ when the selection function predicts that prediction should be given, or abstains if the selection function predicts that it should not predict. Convenient way to formulate the selection function $g$ is relying on a confidence function $\tilde{g}$ and a threshold $\gamma \in \mathbb{R}$ as:

$$g(x) = \mathbb{1}[\tilde{g}(x) > \gamma] \quad (1)$$

where confidence function $\tilde{g} : X \rightarrow \mathbb{R}$ assigns a real-valued confidence to an instance $x \in X$. Ideally, a good confidence estimator $\tilde{g}(x)$ for abstention should yield high values when $f(x)$ is correct and low values when it is incorrect.

**Metrics for Selective Prediction** Coverage ($C$) is the portion of instances that the model choose to predict, while risk ($R$) is the error on that subset of predictions. For a selective classifier $h = (f, g)$ on

dataset $D$ with inputs $x_i$ and ground truth labels $y_i$, they are given as follows:

$$C(h) = \frac{1}{|D|} \sum_{(x_i,y_i)\in D} g(x_i) \tag{2}$$

$$R(h) = \frac{\frac{1}{|D|} \sum\limits_{(x_i,y_i)\in D} l(f(x_i), y_i) g(x_i)}{C(h)} \tag{3}$$

where loss function $l$ measures the error between the predicted label $f(x_i)$ and the ground truth $y_i$. Ideally, a reliable model should showcase high coverage at low levels of risk, implying accurate predictions for many instances and abstention on others. As the threshold $\gamma$ in Eq. 1 decreases, coverage increases, but risk also rises. Hence there exists a risk-coverage trade-off that models strive to optimize. Thus, we construct a curve plotting coverage versus the corresponding risk (El-Yaniv et al., 2010) and calculate the Area Under Risk-Coverage Curve (AURCC). A lower AURCC indicates a better selective classifier.

We calculate Reversed Pair Proportion (RPP) following Xin et al. 2021, a normalized version of the Kendall-Tau distance (Kendall, 1948) to gauge how closely the confidence estimator $\tilde{g}$ aligns with the ideal. Ideally, the confidence estimator should rank all incorrect predictions below all correct predictions. RPP quantifies the proportion of instance pairs with a reversed confidence–error relationship.

$$RPP = \frac{\sum\limits_{1\le i,j\le |D|} \mathbb{1}[\tilde{g}(x_i) < \tilde{g}(x_j), l_i < l_j]}{|D|^2} \tag{4}$$

A lower RPP value indicates better estimator. However, Gu and Hopkins 2023 highlights that RPP and AUC values are influenced by both the prediction and confidence estimation functions. They propose refinement metric, normalizing with the worst-case Kendall-tau distance, offering a calibrated interpretable metric where 0, 0.5 and 1 signifies the best, random and the worst case respectively.

$$Rf = \frac{\sum\limits_{1\le i,j\le |D|} \mathbb{1}[\tilde{g}(x_i) < \tilde{g}(x_j), l_i < l_j]}{c(|D| - c)} \tag{5}$$

where $c$ denote number of correct predictions made by the prediction function.

### 4.1 Confidence Estimators

**Softmax Response (SR)** (Hendrycks and Gimpel, 2016) derives the confidence estimate based on the maximum probability assigned to one of the labels.

$$\tilde{g}_{SR}(x) = \max_{y\in Y} p(y) \tag{6}$$

**Monte Carlo (MC) Dropout** (Gal and Ghahramani, 2016) Different dropout value is used to derive the confidence estimate of a neural network by computing $p(y)$ for a total of N times and we consolidate these N probability values into a confidence estimate using the below three variants.

*Sampled maximum probability* (SMP) uses the sample mean as the final confidence. $p_y^n$ is the probability of the label $y$ obtained using the $n^{th}$ mask.

$$\tilde{g}_{SMP}(x) = \max_{y\in Y} \frac{1}{N} \sum_{n=1}^{N} p^n(y) \tag{7}$$

*Probability Variance* (PV) (Gal and Ghahramani, 2016) computes negative variance of them. $\bar{p}_y$ is mean probability for a label $y$ across N values.

$$\tilde{g}_{PV}(x) = \frac{1}{|Y|} \sum_{y\in Y} \frac{1}{N} \sum_{n=1}^{N} (p_y^n - \bar{p}_y)^2 \tag{8}$$

*Bayesian active learning by disagreement* (BALD) (Houlsby et al., 2011; Vazhentsev et al., 2022) measures using the mutual information as follows:

$$\tilde{g}_{BALD}(x) = \sum_{y\in Y} \bar{p}_y \log \bar{p}_y + \sum_{y,n} p_y^n \log p_y^n \tag{9}$$

This MC dropout mechanism is equivalent to using an ensemble model for confidence estimation, but does not require actually training and storing multiple models but with increased inference cost.

### 4.2 Training Loss

**Confident Error Regularizer** (Xin et al., 2021) adds an additional regularizer, along with task specific loss, aims to optimize for RPP at training time as if the model's error on example exceeds its error on other example (i.e current example is more difficult), then the confidence on that example should not surpass the confidence on other example.

$$L_{CER} = \sum_{1\le i,j\le N} \Delta_{i,j} \mathbb{1}[e_i > e_j] \tag{10}$$

$$\Delta_{i,j} = \max\{0, \max_{y\in Y} p_i(y) - \max_{y\in Y} p_j(y)\}^2 \tag{11}$$

here N is the number of instances in a batch and $e_i$ is an error of the $i^{th}$ instance and use SR to obtain confidence here as it is easily accessible at training time, while MC-dropout confidence is not.

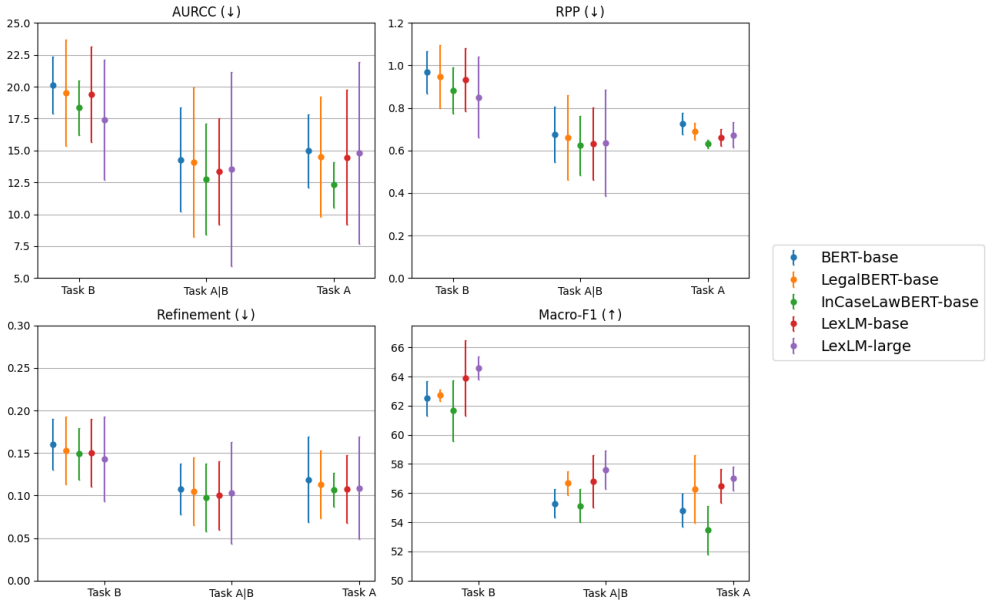


Figure 1: Impact of pre-trained model on selective prediction and classification performance.

**Expected Calibration Error (ECE) Loss** (Shamsi et al.) is added additionally to task-specific loss and is calculated by grouping the predictions into different bins (M bins) according to their confidence. The final loss is aggregated across bins where the error of each bin is computed as the difference between the accuracy and the confidence as:

$$L_{ECE} = \sum_{m \in M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (12)$$

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}_{\hat{y}_i = y_i} \;\&\; \text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} p_i$$

where N represents number of instances, $\hat{y}_i$ and $y_i$ indicate predicted and actual labels respectively.

**Gambler's Loss** (Liu et al., 2019) Unlike above methods which derive abstention label based on confidence estimate of actual label predictions, Gambler's loss explicitly augments an extra class to learn the selection function and trains with the loss function that allows the prediction function to benefit from abstaining on difficult instances:

$$L_{gambler} = \sum_{y \in Y} I(y) \log[p(y) + \frac{1}{r} p(abs)] \quad (13)$$

where $I(y)$ is binary indicator indicating if $y$ is the true label, $p(abs)$ denotes the rejection score and and $r$ is the rejection reward hyperparameter. Here the coverage is obtained by varying the threshold of abstention logit.

### 4.3 Extending to Multi-label case

While all the above techniques are typically proposed for multi-class classification scenarios, we adapt them to the multi-label setting for our COC task by treating each label as a separate binary classification task. This involves deriving a confidence estimate for each label by utilizing the probability assigned to that label and its complement (1 - probability). Similarly, we vary the threshold for each label independently to facilitate abstention and the evaluation metrics are computed for each label separately. We report the macro-average across all the labels for an instance, unless specified.

## 5 Experiments

Following Chalkidis et al. 2022a, we employ a hierarchical extension of a pre-trained transformer model to account for longer input texts as our base model. We use corresponding pre-trained model to encode each paragraph in the input independently to obtain [cls] representation for each paragraph. We then pass these paragraph representations to a transformer layer to learn contextual information from other paragraphs. Finally, we max pool over these context-aware paragraph representations to obtain final representation of case facts which is sent to classification layer. In case of Task A|B, we concatenate a multi-hot feature vector containing the task B labels to the final representation before passing it to the classifier as in Santosh et al. 2022a. Detailed hyperparameters can be found in App. A.

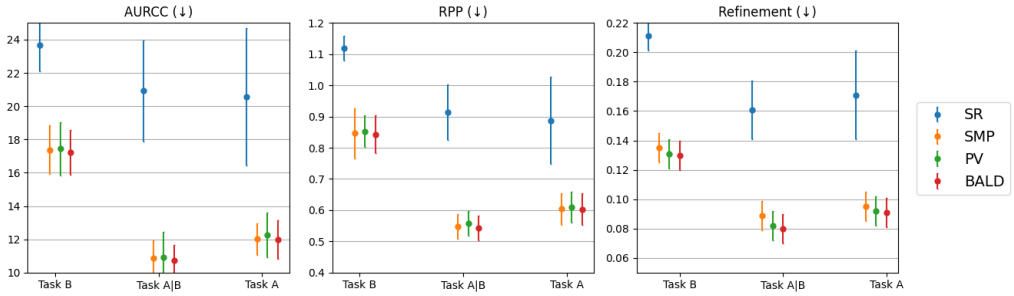


Figure 2: Impact of confidence estimators on selective prediction and classification performance.

To assess the reliability of COC models using diverse pre-trained language models as backbone, we employ (i) BERT-base (Kenton and Toutanova, 2019), trained on general english corpora (ii) InCaseLawBERT-base (Paul et al., 2023), pre-trained on Indian Supreme Court and High Courts case documents by initializing with CaseLawBERT (Zheng et al., 2021) which is pre-trained on US Case Law from federal and state courts. (iii) LegalBERT-base (Chalkidis et al., 2020) incorporates pre-training on EU, UK, and cases from the US, European Court of Justice, and ECtHR. (iv) LexLM-base (Chalkidis et al., 2023), pre-trained on LeXFiles, a diverse legal corpus spanning US, UK, EU, India, and Canada jurisdictions. (v) LexLM-large (Chalkidis et al., 2023), a larger version pre-trained on the same LexFiles corpus. Among them, BERT lacks specific legal pre-training, while InCaseLawBERT has access to legal corpus but not from the ECtHR jurisdiction, relevant to our COC task. LegalBERT and LexLM have access to ECtHR corpus in pre-training, but they differ in the proportion of ECtHR corpus to other corpora, where LexLM has less ECtHR proportion compared to LegalBERT's pre-training corpora.

We use these 5 pre-trained models as backbone in base model and fine-tune on COC task with 4 training loss functions (i.e task-specific loss, CER, ECE, Gambler in Sec. 4.2). We employ 4 variants (SR, SMP, PV, BALD in Sec. 4.1) to derive the confidence estimate and thus compute the selective prediction metrics (AURCC, RPP, Refinement) and classification metric (macro-F1) across these 80 (5*4*4) configurations for each of the three tasks.

### 5.1 Results

**(a) Impact of Pre-trained models:** Fig. 1 reports the selective prediction and classification performance metrics , averaged across all the configurations (confidence estimators, training losses) for each model for three COC tasks along with standard deviation. Lower scores are preferable for selective metrics, while higher scores are desired for macro-F1. Our observations reveal that on macro-F1 scores, BERT consistently outperforms InCaseLawBERT across all three tasks consistently. LegalBERT, LexLM-Base, and LexLM-Large outperform BERT's performance in that order. We hypothesize that InCaseLawBERT, despite being pre-trained on domain-specific (legal) US and Indian contexts, may not serve as a better starting point for generalizing to the ECtHR COC task. In contrast, models with access to the ECtHR corpus, such as LegalBERT, LexLM-Base, and LexLM-Large, benefit from pre-training, with LexLM's diverse corpus contributing to better generalization capabilities.

On examining selective metrics, BERT-base exhibits the highest (worst) score compared to legally pre-trained models, underscoring that domain-specific pre-training, can aid not only in better accuracy but also lead to better calibration than general ones. Among the legal base models, the proportion of ECtHR corpus in pre-training inversely correlates with scores on selective metrics—the model with no access (InCaseLawBERT) performs the best, followed by LexLM and LegalBERT. This trend is consistent across the three selective metrics. The tendency of models with access to ECtHR corpus in pre-training resulting in overconfident predictions on downstream tasks, may be attributed to spurious artifacts present in the data, influencing the model right from the pre-training stage. This pattern persists across all three tasks. Surprisingly, LexLM-Large maintains the best selective performance on Task B despite having access to ECtHR in pre-training. This could be attributed to its higher number of parameters enabling better generalization. However, the effect of greater parameterization in LexLM-Large diminishes when

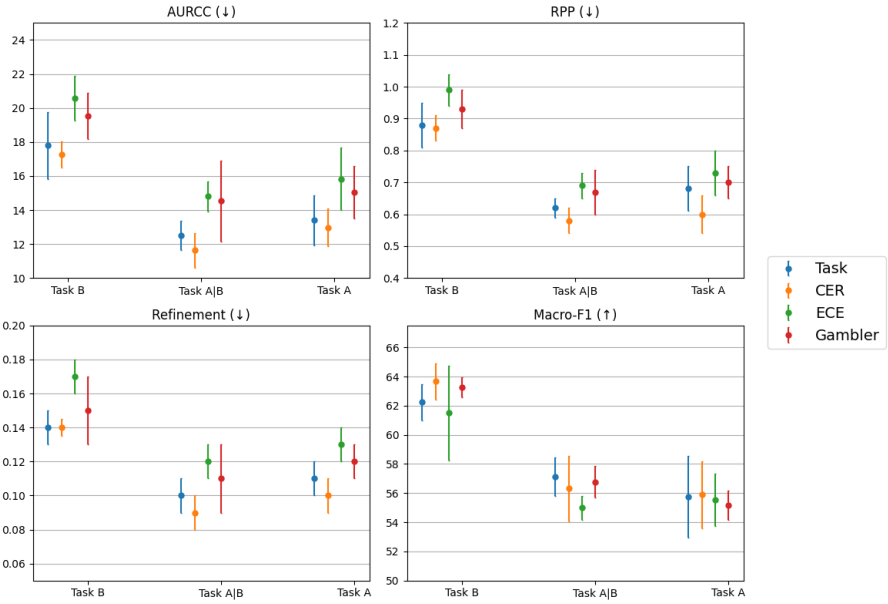


Figure 3: Impact of training loss function on selective prediction and classification performance.

transitioning to more challenging tasks (Task A|B, A), rendering it overconfident.

**Main Takeaways:** Our findings underscore that (i) Access to a domain-specific corpus during pre-training improves model calibration compared to its absence. (ii) However, excessive focus on a downstream domain corpus during pre-training may negatively impact calibration, causing the model to pick up spurious artifacts during pre-training phase that turn challenging to unlearn during fine-tuning. This highlights the need for a diverse domain-related corpus during pre-training to strike a balance between accuracy and confidence. Additionally, the development of effective saliency masking strategies during pre-training is crucial for producing more robust and reliable pre-trained models, steering away from spurious artifacts. (iii) Larger models exhibit overconfidence compared to their base versions, despite achieving higher accuracy.

**(b) Impact of Confidence Estimation** We present averaged selective prediction metrics[1] along with standard deviation across various models and training loss function configurations for each confidence estimator in Fig. 2. We observe that computationally intensive MC dropout variants - SMP, PV, BALD - achieve significant improvements over SR on all metrics and tasks consistently, aligning with Vazhentsev et al. 2022 and contrary to the findings of Xin et al. 2021, where the opposite trend is observed. Gu and Hopkins (2023) advocate to use refinement to compare confidence estimators, in contrast to AURCC and RPP, commonly used in prior works (Xin et al., 2021; Vazhentsev et al., 2022, 2023; Whitehead et al., 2022) as as they are also influenced by the effectiveness of the base prediction function. On refinement among the three MC methods, BALD takes the slightest lead on all tasks, followed by PV and SMP, albeit marginally. The effectiveness of BALD compared to PV and SMP may stem from the latter focusing exclusively on epistemic uncertainty arising from a lack of knowledge, ignoring aleatoric uncertainty associated with ambiguity and noise in the data, while the former measures total uncertainty (Vazhentsev et al., 2022; Malinin and Gales, 2018).

**Main Takeaways:** MC Dropout notably enhances confidence estimation compared to SR, albeit with an increase in computational costs during inference. This raises concerns, considering the substantial computational overhead with these pre-trained models. Therefore, the development of effective yet computationally light confidence estimators represents a potential avenue for further exploration.

**(c) Impact of Training Loss** We report the selective prediction and performance metrics averaged across all configurations for each training loss in Fig. 3. We observe that adding confident error regularizer boosted all the selective metrics consistently across all the tasks than trained with task-

[1]The choice of confidence estimator does not affect the models' performance as it is dependent on the pre-trained model and loss function.

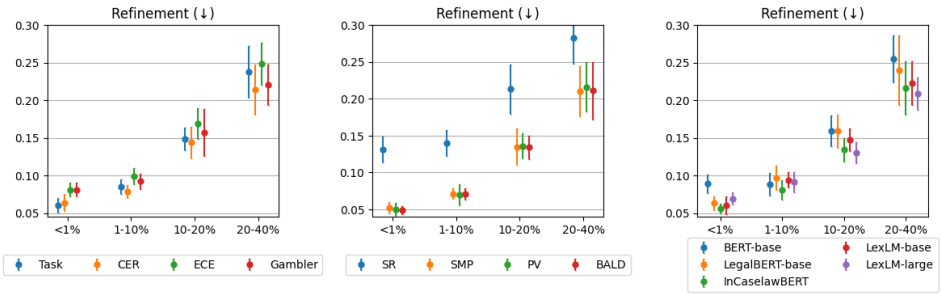


Figure 4: Impact of pre-trained model, confidence estimator and loss function on different labels bucketed by frequency of cases in which they deem to be violated in Task A.

specific loss alone. This can be attributed to its design to deliberately alleviate overconfidence during fine-tuning. Notably, it did not harm model accuracy and maintained comparable or even better to baseline in some cases. On the other hand, ECE regularization negatively impacts both confidence calibration and the performance than normal one. This is attributed to the non-differentiable nature of ECE loss, prompting the need for further investigation into differentiable surrogates of ECE loss, as proposed in (Karandikar et al., 2021; Bohdal et al., 2021). Gamblers loss which allows model to directly learn abstention head, maintained comparable performance to baseline/CER but witnessed a drop in selective prediction performance.

**Main Takeaways:** Adding CER is the most compelling recipe for fine-tuning to balance accuracy with confidence calibration.

**(d) Effect on different labels** To assess the influence of each design factor on different labels, we categorize the labels based on their frequency of occurrence in cases deemed to be violated in the training set (Task A). Thus we obtain four buckets with 5, 4, 4, 1 articles accounting for <1% (rarely violated), 1-10%, 10-20% and 20-40% (frequently violated) respectively. We calculate the averaged refinement scores across articles in each bucket, considering various configurations under different training losses, confidence estimators and pre-trained models and present the results in Fig. 4. With respect to training loss, we observe that adding CER performs better than baseline task specific loss alone, as we move towards more frequent labels. This can also be due to less number of violated cases available for rare articles to effectively regularize them, as the probability of them appearing in same batch tends to be lower. While ECE and Gamblers underperformed compared to task-specific loss in rare article buckets, but gamblers picked it up towards frequent bucket, making it comparable to CER.

Among confidence estimators, MC dropout methods consistently maintained better performance then SR across all the buckets. Across pre-trained models, InCaseLawBERT maintains better performance across all the buckets and trend of decreasing performance with increasing access to ECtHR can also be noticed from LexLM-base and LegalBERT. On the other hand, BERT-base shows the worst performance across all buckets, emphasizing the need of domain-specific pre-training for a better calibrated model. However, LexLM-large suffers in rare article buckets, but picks up in the frequent article buckets due to presence of larger number of parameters to capture diverse signals in frequent cases. Models' overconfidence values increase towards the frequent buckets due to the confounding effect of more positive data instances.

## 6 Conclusion

We introduce the problem of selective prediction for COC task, aiming to enhance model reliability by abstaining in cases of low confidence. Through empirical investigation on 3 COC task variants with 5 pre-trained models, 4 confidence estimators, and 4 loss functions, we assess how these design factors contribute to better selective prediction in COC. Our findings reveal that legal domain-specific pre-trained models outperform in classification-related metrics and are well-calibrated, but an exclusive focus on a specific corpus can prove detrimental. Larger models tend to be overconfident compared to their base versions. Despite being computationally intense, MC Dropout methods provide superior confidence estimates. Adding CER regularization helps alleviate overconfidence. We hope this preliminary investigation spurs additional research into the selective prediction of COC models, em-

phasizing the critical need for models to be aware of what they don't know in high-stakes domains like Legal COC.

## Limitations

Our study is limited by the datasets, models and selective prediction techniques we consider. We rely on the ECtHR dataset and the findings may be influenced by any characterstics present in this particular dataset such as spurious correlations in the downstream task and effectiveness of simple token based decision trees for such tasks (Santosh et al., 2022a). Extending the study to diverse datasets and legal jurisdictions would contribute to a more comprehensive understanding of the reliability of design factors associated with Case Outcome Classification models in legal NLP and thus bolstering the generalizability.

Due to computational constraints, we are unable to pre-train language models from scratch with specific combinations, such as exclusive pre-training on the ECtHR corpus, larger model size variants with billions of parameters, or different pre-training corporus combinations, hindering our ability to conduct dedicated ablations on our claims. Consequently, we rely on existing pre-trained models and solely undertake fine-tuning in this study. While we have made efforts to include a diverse range of selective prediction techniques in our empirical investigation, it is important to acknowledge that our study may not comprehensively cover all the work in this space. Moreover, while we have selected certain design factors for examination, there are other architectural variants proposed in recent studies such as dependency learning across tasks, incorporation of loss functions like contrastive learning, integration of additional external legal knowledge, like legal articles. These aspects warrant exploration in future research endeavors to assess their impact on model reliability.

Additionally, we advocate for future studies in COC to not only report prediction performance but also include reliability metrics to provide deeper insights into models' confidence calibration. This would offer a more comprehensive understanding of model behavior and enhance the trustworthiness of COC models in legal contexts.

## Ethics Statement

Our dataset is derived from LexGLUE benchmark (Chalkidis et al., 2022a) which is obtained from a publicly available database of ECtHR decisions, available in the public court database HUDOC[2]. Despite the inclusion of real names and the absence of anonymization in these decisions, we do not foresee any harm resulting from our experiments beyond the disclosure of this information.

The task of COC/LJP raises significant ethical and legal considerations, both in a broad context and particularly concerning the European Court of Human Rights (Fikfak, 2021). It is crucial to clarify that our research does not advocate for the practical implementation of Case Outcome Classification (COC) within courts. The experimental nature of our study is designed to explore the reliability of models in controlled settings and does not propose or endorse real-world deployment within legal systems. Our results are hence to be understood as technical contributions in pursuit of the overarching goal of developing models capable of deriving insight from data that can be used legally, ethically, and mindfully by experts in solving problems arising in legal research and practice.

We acknowledge that, in adapting pre-trained encoders, our models may inherit existing biases. Similarly, the ECtHR case collection, being historical data, may exhibit a data distribution where sensitive attributes of individuals (e.g., applicant gender) could offer predictive signals for the allegation/violation variable, as demonstrated in previous work (Chalkidis et al., 2022b). Although we believe the results observed in our COC experiments are not substantially connected to such encoded bias, it is crucial to highlight that legal NLP systems utilizing case outcome information and intended for practical deployment should undergo thorough scrutiny against relevant equal treatment imperatives, ensuring scrutiny of their performance, behavior, and intended use.

## References

Mohammad Alali, Shaayan Syed, Mohammed Alsayed, Smit Patel, and Hemanth Bodala. 2021. Justice: A benchmark dataset for supreme court's judgment prediction. *arXiv preprint arXiv:2112.03414*.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

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

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. Interpreting predictive probabilities: Model confidence or human label variation? *arXiv preprint arXiv:2402.16102*.

Nina Baumgartner, Matthias Stürmer, Matthias Grabmair, Joel Niklaus, et al. 2024. Towards explainability and fairness in swiss judgement prediction: Benchmarking on a multilingual dataset. *arXiv preprint arXiv:2402.17013*.

Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. 2021. Meta-calibration: Learning of model calibration using differentiable expected calibration error. *arXiv preprint arXiv:2106.09613*.

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. *arXiv preprint arXiv:2103.13084*.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. *arXiv preprint arXiv:2305.07507*.

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.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Felix Schwemer, and Anders Søgaard. 2022b. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. *arXiv preprint arXiv:2203.07228*.

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan O Arik, Tomas Pfister, and Somesh Jha. 2023. Adaptation with self-evaluation to improve selective prediction in llms. *arXiv preprint arXiv:2310.11689*.

Chi-Keung Chow. 1957. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254.

Jeremy R Cole, Michael JQ Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. 2016. Learning with rejection. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 67–82. Springer.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302.

Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. Juribert: A masked-language model adaptation for french legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101.

Ran El-Yaniv et al. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5).

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

Giorgio Fumera and Fabio Roli. 2002. Support vector machines with embedded reject option. In *Pattern Recognition with Support Vector Machines: First International Workshop, SVM 2002 Niagara Falls, Canada, August 10, 2002 Proceedings*, pages 68–82. Springer.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Leilei Gan, Kun Kuang, Yi Yang, and Fei Wu. 2021. Judgment prediction via injecting legal knowledge into neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12866–12874.

Siddhant Garg and Alessandro Moschitti. 2021. Will this question be answered? question filtering via answer model distillation for efficient question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7329–7346.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Zhengyao Gu and Mark Hopkins. 2023. On the evaluation of neural selective prediction methods for natural language processing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7899.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Martin E Hellman. 1970. The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics*, 6(3):179–185.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *stat*, 1050:24.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.

Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. 2018. To trust or not to trust a classifier. *Advances in neural information processing systems*, 31.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696.

Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Becca Roelofs. 2021. Soft calibration objectives for neural networks. *Advances in Neural Information Processing Systems*, 34:29768–29779.

Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, 12(4):e0174698.

Aaron Russell Kaufman, Peter Kraft, and Maya Sen. 2019. Improving supreme court forecasting using boosted decision trees. *Political Analysis*, 27(3):381–387.

Maurice George Kendall. 1948. Rank correlation methods.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Fred Kort. 1957. Predicting supreme court decisions mathematically: A quantitative analysis of the "right to counsel" cases. *American Political Science Review*, 51(1):1–12.

Kankawin Kowsrihawat, Peerapon Vateekul, and Prachya Boonkwan. 2018. Predicting judicial decisions of criminal cases from thai supreme court using bi-directional gru with attention mechanism. In *2018 5th Asian Conference on Defense Technology (ACDT)*, pages 50–55. IEEE.

André Lage-Freitas, Héctor Allende-Cid, Orivaldo Santana, and Lívia Oliveira-Lage. 2022. Predicting brazilian court decisions. *PeerJ Computer Science*, 8:e904.

Yanjun Li, Huan Huang, Qiang Geng, Xinwei Guo, and Yuyu Yuan. 2022. Fairness measures of machine learning models in judicial penalty prediction. *Journal of Internet Technology*, 23(5):1109–1116.

Yifei Liu, Yiquan Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023. Mlljp: Multi-law aware legal judgment prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1023–1034.

Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. 2019. Deep gamblers: Learning to abstain with portfolio theory. *Advances in Neural Information Processing Systems*, 32.

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.

Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31.

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*, pages 12–23.

Stuart S Nagel. 1963. Applying correlation analysis to case prediction. *Tex. L. Rev.*, 42:1006.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35.

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2023. Pre-trained language models for the legal domain: a case study on indian law. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 187–196.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2022. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

T. Y. S. S Santosh, Oana Ichim, and Matthias Grabmair. 2023a. Zero shot transfer of article-aware legal outcome classification for european court of human rights cases. *arXiv preprint arXiv:2302.00609*.

TYS Santosh, Marcel Perez San Blas, Phillip Kemper, and Matthias Grabmair. 2023b. Leveraging task dependency and contrastive learning for case outcome classification on european court of human rights cases. *arXiv preprint arXiv:2302.00768*.

TYS Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022a. Deconfounding legal judgment prediction for european court of human rights cases towards better alignment with experts. *arXiv preprint arXiv:2210.13836*.

T.y.s.s Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022b. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeffrey A Segal. 1984. Predicting supreme court cases probabilistically: The search and seizure cases, 1962-1981. *American Political Science Review*, 78(4):891–900.

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.

Afshar Shamsi, Hamzeh Asgharnezhad, AmirReza Tajally, Saeid Nahavandi, and Henry Leung. An uncertainty-aware loss function for training neural networks with calibrated predictions.

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.

Octavia-Maria Şulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef van Genabith. 2017a. Exploring the use of text classification in the legal domain.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017b. Predicting the law area and decisions of french supreme court cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722.

Santosh Tyss, Oana Ichim, and Matthias Grabmair. 2023a. Zero-shot transfer of article-aware legal outcome classification for european court of human rights cases. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 593–605.

Santosh Tyss, Marcel Perez San Blas, Phillip Kemper, and Matthias Grabmair. 2023b. Leveraging task dependency and contrastive learning for case outcome classification on european court of human rights cases. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1103–1103.

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.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022a. Investigating selective prediction approaches across several tasks in iid, ood, and adversarial settings. In *60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, pages 1995–2002. Association for Computational Linguistics (ACL).

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022b. Towards improving selective prediction ability of nlp systems. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 221–226.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, et al. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252.

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681.

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.

Bernhard Waltl, Georg Bonczek, Elena Scepankova, Jörg Landthaler, and Florian Matthes. 2017. Predicting the outcome of appeal decisions in germany's tax law. In *International conference on electronic participation*, pages 89–99. Springer.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. *arXiv preprint arXiv:2005.00963*.

Yuzhong Wang, Chaojun Xiao, Shirong Ma, Haoxi Zhong, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2021. Equality before the law: Legal judgment consistency analysis for fairness. *arXiv preprint arXiv:2103.13868*.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. 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 1040–1051.

Shanshan Xu, Oana Ichim, Isabella Risini, Barbara Plank, Matthias Grabmair, et al. 2023. From dissonance to insights: Dissecting disagreements in rationale construction for case outcome classification. *arXiv preprint arXiv:2310.11878*.

Shanshan Xu, TYS Santosh, Oana Ichim, Barbara Plank, and Matthias Grabmair. 2024. Through the lens of split vote: Exploring disagreement, difficulty and calibration in legal case outcome classification. *arXiv preprint arXiv:2402.07214*.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4085–4091.

Hiyori Yoshikawa and Naoaki Okazaki. 2023. Selective-lama: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1972–1983.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 973–982.

Han Zhang, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2023. Contrastive learning for legal judgment prediction. *ACM Transactions on Information Systems*, 41(4):1–25.

Lili Zhao, Linan Yue, Yanqing An, Yuren Zhang, Jun Yu, Qi Liu, and Enhong Chen. 2022. Cpee: C ivil case judgment p rediction centering on the trial mode of e ssential e lements. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2691–2700.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

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.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1250–1257.

## A Implementation Details

We utilize the Adam optimizer (Kingma and Ba, 2014) to train our models, starting with an initial learning rate of 3e-5 for base and 3e-6 for larger models. Early stopping on validation data for up to 20 epochs is also employed. To reduce memory usage during training, we use mixed precision (fp16) and gradient accumulation. These models can handle 64 paragraphs, each with 128 tokens. The batch size is set to 8 in all base and 2 in large experiments. For MC Dropout methods, we use 10 runs. For CER, we add the regularization loss with a weight within $\{0.01, 0.05, 0.1, 0.5\}$ to choose the best one on AURCC metric. For ECE, we restrict to 10 bins. For Gamblers loss, we vary the reward within $\{1.0, 5.0, 6.5, 14.0\}$ with 4 warm-up epochs. We conduct each experiment five times with different random seeds and calculate their mean values.

## B Detailed Experimental Results

We provide the selective prediction and classification performance metrics for Task B in Tables 1, 2, 3, Task A|B in Tables 5, 6, 4, and Task A in Tables 7, 8, 9. Across all configurations, MC Dropout methods consistently outperformed SR, with BALD leading or remaining comparable in most settings. However, in specific configurations such PV on LexLM-Base with CER, PV outperformed BALD.

On training loss choice, CER demonstrated improvements over task-specific loss alone, while Gambler's loss showed competitiveness in specific configurations, surpassing even CER in instances like InCaseLawBERT-base for Task A and Task B, as well as LexLM-large for Task A|B. Interestingly, ECE lagged in most cases but outperformed CER in InCaseLawBERT-base for Task B and A|B, as well as LexLM-large for Task A|B, aligning with Gambler's superior performance in those configurations. These findings suggest a nuanced interplay between these specific models and data characteristics, prompting further investigation into their interactions and a theoretical understanding of their properties. Regarding pre-trained models, LexLM-base, particularly with certain configurations like PV and CER on Task A and B, attempted to yield better results than LexLM-large or InCaseLawBERT-base, warranting further exploration into the interplay between these techniques.

| Confidence Estimator | Models Loss | BERT-base AURCC | RPP | Rf | mac.-F1 | LegalBERT-base AURCC | RPP | Rf | mac.-F1 |
|---|---|---|---|---|---|---|---|---|---|
| SR | Task | 21.94 | 1.055 | 0.186 | 61.34 | 28.42 | 1.055 | 0.173 | 62.42 |
| SMP | Task | 18.82 | 0.915 | 0.139 | 61.34 | 16.49 | 1.124 | 0.110 | 62.42 |
| PV | Task | 21.48 | 0.909 | 0.135 | 61.34 | 16.36 | 0.830 | 0.111 | 62.42 |
| BALD | Task | 17.98 | 0.905 | 0.134 | 61.34 | 16.22 | 0.831 | 0.110 | 62.42 |
| SR | CER | 22.19 | 1.099 | 0.209 | 62.60 | 21.11 | 1.035 | 0.201 | 62.31 |
| SMP | CER | 18.15 | 0.883 | 0.135 | 62.60 | 16.26 | 0.784 | 0.125 | 62.31 |
| PV | CER | 18.05 | 0.882 | 0.132 | 62.60 | 16.30 | 0.803 | 0.127 | 62.31 |
| BALD | CER | 17.92 | 0.873 | 0.129 | 62.60 | 16.04 | 0.785 | 0.126 | 62.31 |
| SR | ECE | 24.19 | 1.165 | 0.204 | 61.75 | 27.67 | 1.283 | 0.244 | 63.32 |
| SMP | ECE | 19.30 | 0.925 | 0.150 | 61.75 | 19.19 | 0.919 | 0.217 | 63.32 |
| PV | ECE | 19.09 | 0.926 | 0.149 | 61.75 | 19.36 | 0.936 | 0.141 | 63.32 |
| BALD | ECE | 18.96 | 0.926 | 0.147 | 61.75 | 19.26 | 0.921 | 0.141 | 63.32 |
| SR | Gamb | 25.21 | 1.185 | 0.246 | 64.32 | 26.16 | 1.223 | 0.226 | 62.80 |
| SMP | Gamb | 19.58 | 0.941 | 0.152 | 64.32 | 17.91 | 0.867 | 0.132 | 62.80 |
| PV | Gamb | 19.75 | 0.959 | 0.158 | 64.32 | 17.94 | 0.884 | 0.132 | 62.80 |
| BALD | Gamb | 19.64 | 0.948 | 0.163 | 64.32 | 17.80 | 0.872 | 0.131 | 62.80 |

Table 1: Task B

| Confidence Estimator | Models Loss | LexLM-base AURCC | RPP | Rf | mac.-F1 | LexLM-large AURCC | RPP | Rf | mac.-F1 |
|---|---|---|---|---|---|---|---|---|---|
| SR | Task | 19.33 | 0.962 | 0.199 | 61.42 | 17.92 | 0.914 | 0.177 | 64.31 |
| SMP | Task | 16.86 | 0.835 | 0.129 | 61.42 | 14.06 | 0.729 | 0.120 | 64.31 |
| PV | Task | 16.77 | 0.840 | 0.129 | 61.42 | 14.04 | 0.734 | 0.118 | 64.31 |
| BALD | Task | 16.56 | 0.833 | 0.129 | 61.42 | 14.42 | 0.745 | 0.120 | 64.31 |
| SR | CER | 20.29 | 1.092 | 0.198 | 65.21 | 25.65 | 1.178 | 0.222 | 64.57 |
| SMP | CER | 16.85 | 0.801 | 0.122 | 65.21 | 14.72 | 0.660 | 0.112 | 64.57 |
| PV | CER | 15.37 | 0.810 | 0.121 | 65.21 | 14.87 | 0.759 | 0.112 | 64.57 |
| BALD | CER | 17.06 | 0.814 | 0.121 | 65.21 | 14.50 | 0.698 | 0.109 | 64.57 |
| SR | ECE | 29.07 | 1.336 | 0.246 | 58.24 | 28.78 | 1.293 | 0.280 | 65.82 |
| SMP | ECE | 20.36 | 0.963 | 0.145 | 58.24 | 16.18 | 0.824 | 0.131 | 65.82 |
| PV | ECE | 20.29 | 0.972 | 0.143 | 58.24 | 16.18 | 0.838 | 0.131 | 65.82 |
| BALD | ECE | 19.95 | 0.950 | 0.141 | 58.24 | 16.08 | 0.836 | 0.130 | 65.82 |
| SR | Gamb | 27.06 | 1.171 | 0.211 | 62.70 | 25.37 | 1.163 | 0.156 | 63.65 |
| SMP | Gamb | 18.36 | 0.854 | 0.124 | 62.70 | 15.09 | 0.738 | 0.124 | 63.65 |
| PV | Gamb | 18.36 | 0.856 | 0.122 | 62.70 | 15.26 | 0.752 | 0.124 | 63.65 |
| BALD | Gamb | 17.87 | 0.846 | 0.119 | 62.70 | 15.25 | 0.740 | 0.122 | 63.65 |

Table 2: Task B

| Confidence Estimator | Models Loss | InCaseLawBERT-base AURCC | RPP | Rf | mac.-F1 |
|---|---|---|---|---|---|
| SR | Task | 19.29 | 0.961 | 0.172 | 61.70 |
| SMP | Task | 16.22 | 0.801 | 0.122 | 61.70 |
| PV | Task | 16.31 | 0.810 | 0.121 | 61.70 |
| BALD | Task | 16.27 | 0.801 | 0.119 | 61.70 |
| SR | CER | 21.11 | 1.038 | 0.194 | 63.69 |
| SMP | CER | 17.63 | 0.678 | 0.132 | 63.69 |
| PV | CER | 17.61 | 0.842 | 0.130 | 63.69 |
| BALD | CER | 17.55 | 0.835 | 0.128 | 63.69 |
| SR | ECE | 23.98 | 1.134 | 0.208 | 58.39 |
| SMP | ECE | 18.05 | 0.872 | 0.139 | 58.39 |
| PV | ECE | 17.85 | 0.878 | 0.140 | 58.39 |
| BALD | ECE | 17.62 | 0.879 | 0.141 | 58.39 |
| SR | Gamb | 21.75 | 1.059 | 0.203 | 62.86 |
| SMP | Gamb | 17.44 | 0.834 | 0.141 | 62.86 |
| PV | Gamb | 17.61 | 0.850 | 0.142 | 62.86 |
| BALD | Gamb | 17.52 | 0.842 | 0.140 | 62.86 |

Table 3: Task B

| Confidence Estimator | Model Loss | InCaselawBERT-base AURCC | RPP | Rf | mac.-F1 |
|---|---|---|---|---|---|
| SR | Task | 16.37 | 0.737 | 0.125 | 55.99 |
| SMP | Task | 9.64 | 0.530 | 0.078 | 55.99 |
| PV | Task | 9.79 | 0.543 | 0.079 | 55.99 |
| BALD | Task | 9.71 | 0.535 | 0.078 | 55.99 |
| SR | CER | 12.93 | 0.665 | 0.105 | 54.38 |
| SMP | CER | 9.00 | 0.475 | 0.073 | 54.38 |
| PV | CER | 9.40 | 0.498 | 0.075 | 54.38 |
| BALD | CER | 9.81 | 0.524 | 0.026 | 54.38 |
| SR | ECE | 21.10 | 0.864 | 0.155 | 53.79 |
| SMP | ECE | 10.90 | 0.551 | 0.092 | 53.79 |
| PV | ECE | 10.97 | 0.554 | 0.092 | 53.79 |
| BALD | ECE | 10.83 | 0.562 | 0.093 | 53.79 |
| SR | Gamb | 24.04 | 1.003 | 0.190 | 56.39 |
| SMP | Gamb | 13.39 | 0.649 | 0.103 | 56.39 |
| PV | Gamb | 13.16 | 0.642 | 0.102 | 56.39 |
| BALD | Gamb | 12.84 | 0.645 | 0.103 | 56.39 |

Table 4: Task A|B

| Confidence Estimator | Model Loss | BERT-base AURCC | RPP | Rf | mac.-F1 | LegalBERT-base AURCC | RPP | Rf | mac.-F1 |
|---|---|---|---|---|---|---|---|---|---|
| SR | Task | 15.35 | 0.765 | 0.123 | 56.60 | 18.01 | 0.878 | 0.126 | 57.97 |
| SMP | Task | 11.39 | 0.565 | 0.083 | 56.60 | 11.79 | 0.579 | 0.133 | 57.97 |
| PV | Task | 11.33 | 0.572 | 0.082 | 56.60 | 11.62 | 0.584 | 0.077 | 57.97 |
| BALD | Task | 11.23 | 0.570 | 0.081 | 56.60 | 11.45 | 0.581 | 0.076 | 57.97 |
| SR | CER | 17.49 | 0.817 | 0.134 | 54.09 | 15.01 | 0.750 | 0.110 | 56.36 |
| SMP | CER | 11.74 | 0.578 | 0.079 | 54.09 | 10.10 | 0.512 | 0.069 | 56.36 |
| PV | CER | 11.76 | 0.587 | 0.078 | 54.09 | 10.15 | 0.525 | 0.069 | 56.36 |
| BALD | CER | 11.63 | 0.584 | 0.076 | 54.09 | 10.21 | 0.367 | 0.069 | 56.36 |
| SR | ECE | 20.27 | 0.933 | 0.188 | 54.76 | 27.61 | 1.123 | 0.190 | 55.81 |
| SMP | ECE | 12.12 | 0.608 | 0.112 | 54.76 | 11.91 | 0.614 | 0.086 | 55.81 |
| PV | ECE | 15.31 | 0.608 | 0.112 | 54.76 | 12.05 | 0.624 | 0.086 | 55.81 |
| BALD | ECE | 12.00 | 0.599 | 0.110 | 54.76 | 12.10 | 0.622 | 0.086 | 55.81 |
| SR | Gamb | 26.17 | 0.986 | 0.171 | 55.68 | 28.86 | 1.090 | 0.225 | 56.63 |
| SMP | Gamb | 14.17 | 0.631 | 0.102 | 55.68 | 11.52 | 0.573 | 0.094 | 56.63 |
| PV | Gamb | 14.13 | 0.631 | 0.102 | 55.68 | 11.48 | 0.576 | 0.092 | 56.63 |
| BALD | Gamb | 12.56 | 0.621 | 0.101 | 55.68 | 11.30 | 0.579 | 0.091 | 56.63 |

Table 5: Task A|B

| Confidence Estimator | Model Loss | LexLM-base AURCC | RPP | Rf | mac.-F1 | LexLM-Large AURCC | RPP | Rf | mac.-F1 |
|---|---|---|---|---|---|---|---|---|---|
| SR | Task | 16.26 | 0.818 | 0.134 | 56.06 | 22.56 | 0.987 | 0.162 | 59.00 |
| SMP | Task | 10.86 | 0.543 | 0.155 | 56.06 | 10.36 | 0.510 | 0.069 | 59.00 |
| PV | Task | 10.81 | 0.553 | 0.079 | 56.06 | 10.47 | 0.522 | 0.072 | 59.00 |
| BALD | Task | 10.56 | 0.538 | 0.078 | 56.06 | 10.57 | 0.532 | 0.075 | 59.00 |
| SR | CER | 17.24 | 0.852 | 0.153 | 59.64 | 20.86 | 0.910 | 0.173 | 57.10 |
| SMP | CER | 9.57 | 0.503 | 0.076 | 59.64 | 8.76 | 0.477 | 0.066 | 57.10 |
| PV | CER | 9.65 | 0.510 | 0.077 | 59.64 | 8.83 | 0.488 | 0.067 | 57.10 |
| BALD | CER | 9.76 | 0.517 | 0.078 | 59.64 | 8.75 | 0.394 | 0.066 | 57.10 |
| SR | ECE | 25.22 | 1.102 | 0.189 | 54.94 | 31.39 | 1.207 | 0.237 | 55.71 |
| SMP | ECE | 11.24 | 0.563 | 0.081 | 54.94 | 10.39 | 0.510 | 0.073 | 55.71 |
| PV | ECE | 11.55 | 0.587 | 0.082 | 54.94 | 7.93 | 0.523 | 0.075 | 55.71 |
| BALD | ECE | 11.39 | 0.570 | 0.080 | 54.94 | 9.97 | 0.518 | 0.075 | 55.71 |
| SR | Gamb | 14.41 | 0.717 | 0.112 | 56.55 | 27.29 | 1.071 | 0.210 | 58.60 |
| SMP | Gamb | 10.02 | 0.528 | 0.077 | 56.55 | 8.65 | 0.483 | 0.075 | 58.60 |
| PV | Gamb | 9.81 | 0.525 | 0.076 | 56.55 | 8.59 | 0.488 | 0.075 | 58.60 |
| BALD | Gamb | 9.61 | 0.518 | 0.074 | 56.55 | 8.53 | 0.477 | 0.073 | 58.60 |

Table 6: Task A|B

| Confidence Estimator | Model Loss | BERT-base AURCC | BERT-base RPP | BERT-base Rf | BERT-base mac.-F1 | LegalBERT-base AURCC | LegalBERT-base RPP | LegalBERT-base Rf | LegalBERT-base mac.-F1 |
|---|---|---|---|---|---|---|---|---|---|
| SR | Task | 19.79 | 0.929 | 0.149 | 53.67 | 15.43 | 0.774 | 0.148 | 59.49 |
| SMP | Task | 14.28 | 0.688 | 0.102 | 53.67 | 10.37 | 0.529 | 0.129 | 59.49 |
| PV | Task | 14.14 | 0.693 | 0.103 | 53.67 | 10.24 | 0.530 | 0.079 | 59.49 |
| BALD | Task | 13.75 | 0.670 | 0.102 | 53.67 | 10.13 | 0.525 | 0.077 | 59.49 |
| SR | CER | 17.01 | 0.788 | 0.133 | 53.88 | 15.62 | 0.783 | 0.130 | 57.29 |
| SMP | CER | 13.23 | 0.609 | 0.095 | 53.88 | 11.09 | 0.554 | 0.079 | 57.29 |
| PV | CER | 13.37 | 0.616 | 0.094 | 53.88 | 10.98 | 0.555 | 0.077 | 57.29 |
| BALD | CER | 13.34 | 0.614 | 0.094 | 53.88 | 10.95 | 0.558 | 0.077 | 57.29 |
| SR | ECE | 21.47 | 1.019 | 0.269 | 56.24 | 24.33 | 1.112 | 0.233 | 54.37 |
| SMP | ECE | 14.12 | 0.721 | 0.112 | 56.24 | 13.37 | 0.675 | 0.097 | 54.37 |
| PV | ECE | 14.00 | 0.713 | 0.111 | 56.24 | 20.87 | 0.674 | 0.094 | 54.37 |
| BALD | ECE | 13.93 | 0.709 | 0.110 | 56.24 | 13.00 | 0.660 | 0.092 | 54.37 |
| SR | Gamb | 19.47 | 0.937 | 0.193 | 55.50 | 24.41 | 1.071 | 0.193 | 54.01 |
| SMP | Gamb | 12.61 | 0.642 | 0.111 | 55.50 | 13.80 | 0.672 | 0.103 | 54.01 |
| PV | Gamb | 12.46 | 0.641 | 0.110 | 55.50 | 13.84 | 0.685 | 0.102 | 54.01 |
| BALD | Gamb | 12.45 | 0.638 | 0.110 | 55.50 | 13.56 | 0.666 | 0.100 | 54.01 |

Table 7: Task A

| Confidence Estimator | Model Loss | LexLM-base AURCC | LexLM-base RPP | LexLM-base Rf | LexLM-base mac.-F1 | LexLM-large AURCC | LexLM-large RPP | LexLM-large Rf | LexLM-large mac.-F1 |
|---|---|---|---|---|---|---|---|---|---|
| SR | Task | 19.96 | 0.931 | 0.131 | 56.80 | 22.83 | 0.987 | 0.186 | 56.37 |
| SMP | Task | 8.75 | 0.682 | 0.089 | 56.80 | 10.42 | 0.546 | 0.076 | 56.37 |
| PV | Task | 9.97 | 0.692 | 0.089 | 56.80 | 12.40 | 0.563 | 0.080 | 56.37 |
| BALD | Task | 14.45 | 0.683 | 0.088 | 56.80 | 10.32 | 0.546 | 0.078 | 56.37 |
| SR | CER | 15.83 | 0.287 | 0.147 | 58.20 | 25.98 | 0.955 | 0.208 | 57.08 |
| SMP | CER | 11.44 | 0.577 | 0.078 | 58.20 | 11.54 | 0.474 | 0.063 | 57.08 |
| PV | CER | 8.21 | 0.580 | 0.077 | 58.20 | 9.27 | 0.489 | 0.065 | 57.08 |
| BALD | CER | 11.34 | 0.580 | 0.077 | 58.20 | 9.16 | 0.483 | 0.064 | 57.08 |
| SR | ECE | 23.98 | 0.640 | 0.180 | 55.26 | 27.33 | 1.270 | 0.203 | 58.24 |
| SMP | ECE | 13.73 | 0.678 | 0.115 | 55.26 | 12.33 | 0.602 | 0.097 | 58.24 |
| PV | ECE | 13.82 | 0.687 | 0.112 | 55.26 | 12.54 | 0.619 | 0.096 | 58.24 |
| BALD | ECE | 13.76 | 0.674 | 0.109 | 55.26 | 12.45 | 0.610 | 0.095 | 58.24 |
| SR | Gamb | 27.86 | 1.092 | 0.186 | 55.73 | 29.70 | 1.120 | 0.222 | 56.30 |
| SMP | Gamb | 13.20 | 0.591 | 0.084 | 55.73 | 10.55 | 0.498 | 0.069 | 56.30 |
| PV | Gamb | 13.24 | 0.601 | 0.084 | 55.73 | 10.55 | 0.507 | 0.070 | 56.30 |
| BALD | Gamb | 11.98 | 0.585 | 0.083 | 55.73 | 9.90 | 0.503 | 0.069 | 56.30 |

Table 8: Task A

| Confidence Estimator | Model Loss | InCaseLawBERT-base AURCC | RPP | Rf | mac.-F1 |
|---|---|---|---|---|---|
| SR | Task | 15.04 | 0.763 | 0.125 | 52.43 |
| SMP | Task | 11.60 | 0.588 | 0.090 | 52.43 |
| PV | Task | 11.91 | 0.606 | 0.092 | 52.43 |
| BALD | Task | 12.15 | 0.610 | 0.092 | 52.43 |
| SR | CER | 15.31 | 0.760 | 0.147 | 54.82 |
| SMP | CER | 12.01 | 0.612 | 0.112 | 54.82 |
| PV | CER | 12.02 | 0.617 | 0.111 | 54.82 |
| BALD | CER | 11.96 | 0.610 | 0.110 | 54.82 |
| SR | ECE | 15.09 | 0.764 | 0.111 | 51.38 |
| SMP | ECE | 11.30 | 0.576 | 0.100 | 51.38 |
| PV | ECE | 11.27 | 0.577 | 0.100 | 51.38 |
| BALD | ECE | 11.26 | 0.582 | 0.100 | 51.38 |
| SR | Gamb | 14.90 | 0.757 | 0.134 | 55.24 |
| SMP | Gamb | 10.43 | 0.562 | 0.097 | 55.24 |
| PV | Gamb | 10.35 | 0.564 | 0.096 | 55.24 |
| BALD | Gamb | 10.24 | 0.556 | 0.096 | 55.24 |

Table 9: Task A