

# Through the Lens of Split Vote: Exploring Disagreement, Difficulty and Calibration in Legal Case Outcome Classification

Shanshan Xu<sup>1,2\*</sup>, Santosh T.Y.S.S<sup>1</sup>, Oana Ichim<sup>3</sup>

Barbara Plank<sup>4,5</sup>, Matthias Grabmair<sup>1</sup>

<sup>1</sup>Technical University of Munich, Germany <sup>2</sup>ELTEIMATE

<sup>3</sup>Graduate Institute of International and Development Studies, Switzerland

<sup>4</sup>IT University of Copenhagen, Denmark

<sup>5</sup>LMU Munich & Munich Center for Machine Learning (MCML), Germany

## Abstract

In legal decisions, split votes (SV) occur when judges cannot reach a unanimous decision, posing a difficulty for lawyers who must navigate diverse legal arguments and opinions. In high-stakes domains, understanding the alignment of perceived difficulty between humans and AI systems is crucial to build trust. However, existing NLP calibration methods focus on a classifier’s awareness of predictive performance, measured against the human majority class, overlooking inherent human label variation (HLV). This paper explores split votes as naturally observable human disagreement and value pluralism. We collect judges’ vote distributions from the European Court of Human Rights (ECHR), and present **SV-ECHR**<sup>1</sup> a case outcome classification (COC) dataset with SV information. We build a taxonomy of disagreement with SV-specific subcategories. We further assess the alignment of perceived difficulty between models and humans, as well as confidence- and human-calibration of COC models. We observe limited alignment with the judge vote distribution. To our knowledge, this is the first systematic exploration of calibration to human judgements in legal NLP. Our study underscores the necessity for further research on measuring and enhancing model calibration considering HLV in legal decision tasks.

## 1 Introduction

The task of Case Outcome Classification (COC) involves the classification of legal case outcomes based on textual descriptions of their facts. While achieving high performance is desirable, in high-stakes domains such as legal and medical decisions, the quantification of a model’s predictive confidence, or conversely, its uncertainty, is particularly valuable. It allows experts to make more informed

<sup>1</sup>Our dataset and code is available at [https://github.com/TUMLegalTech/SplitVote\\_ECHR](https://github.com/TUMLegalTech/SplitVote_ECHR)

\*Work done during internship at ELTEIMATE

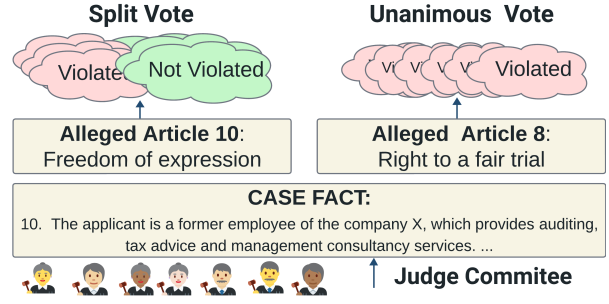


Figure 1: Split Votes in ECtHR Decisions

decisions, especially when the model may be uncertain or where the consequences of a misdiagnosis are significant. Evaluating whether models are aware of their limitations is known as assessing their uncertainty, with popular methods such as *calibration* (Guo et al., 2017; Desai and Durrett, 2020). Calibration assesses the extent to which predictive probabilities accurately reflect the likelihood of a prediction being correct. Models can opt to abstain when the uncertainty exceeds a predefined threshold — a method commonly referred to as *selective classification* (El-Yaniv et al., 2010; Geifman and El-Yaniv, 2017).

Current NLP research focuses on prediction confidence and calibration to assess a classifier’s awareness of its predictive performance only. This evaluation is commonly conducted against the human majority class. However, recent developments in NLP research have shed light on the prevalence of inherent human label variation (HLV) (Plank, 2022), observing disagreement\* across various tasks (Uma et al., 2021). Scholars in the field argue for the acknowledgment and acceptance of HLV, as it mirrors the diverse and pluralistic nature of human values (Sorensen et al., 2024). Notably, Baan et al. 2022 has demonstrated that widely-used calibration metrics may not be suitable for datasets

\*The term HLV embraces disagreement and plausible variation. In this paper, we use these terms interchangeably.

characterized by inherent human disagreement.

In light of these findings, we posit that models should not only be calibrated to recognize their own uncertainty (performance calibration) but also be equipped to discern instances where humans exhibit uncertainty (human calibration) (Baan et al., 2022). This dual focus aims to foster trust among end-users and mitigates potential harm caused by models. Consequently, a critical aspect of this trust involves ensuring the alignment of perceived difficulty between human and models.

This motivates us to study split votes (SV) in court decisions (Fig 1). The judge vote ratio is a *naturally occurring human disagreement* at the European Court of Human Rights (ECtHR). We present SV-ECHR, a COC dataset with judge split vote information. We study the disagreement sources among judges from a legal, linguistic, and NLP perspective. We adapt and analyse the task-agnostic taxonomy components of Xu et al. 2023b and introduce SV-specific subcategories. We also quantitatively assess the effects of different subcategories on judges’ agreement using proxy variables. The results suggest that disagreements are mainly due to the social-political context of cases.

In addition, we assess the alignment of perceived difficulty between COC models and humans, as well as confidence- and human-calibration. While we see acceptable performance in perceived difficulty and confidence calibration, our analysis indicates suboptimal human calibration. This underscores the necessity for a more in-depth inquiry into methods to better align models’ calibration with human behaviour, highlighting opportunities for further research in this direction.

## 2 Related Work

**COC** has been referred to as *Legal Judgment Prediction* (LJP) in previous research (Medvedeva et al., 2020; T.y.s.s et al., 2023a). There exist numerous works involving corpora from ECtHR (Aletras et al., 2016; Chalkidis et al., 2019; T.y.s.s et al., 2023b). All of these approaches mainly focus on the COC performance of models, which is commonly measured against the human majority class. Our SV-ECtHR dataset extends this work and contains the nuanced judges’ vote information. We proceed to use it to systematically investigate disagreement among judges and the alignment of perceived difficulty between models and judges.

**Disagreement / Human Label Variation** is receiving growing attention in mainstream NLP. Various works highlight the presence of HLTV, emphasizing the abundance and plausibility of such human disagreements (Pavlick and Kwiatkowski, 2019; Plank, 2022). Researchers advocate for embracing HLTV for more trustworthy AI (Talat et al., 2022; Casper et al., 2023). As real-world applications are used to assist diverse audiences, it becomes crucial to investigate and include pluralistic human values in NLP systems (Sorensen et al., 2023). This motivates our study of disagreements among judges’ split-votes in the legal decision process. Recently, several studies have proposed task-specific taxonomies to identify potential sources of disagreement in various NLP tasks (Uma et al., 2021; Sandri et al., 2023; Jiang and Marneffe, 2022). Building on the meta-analysis of these existing taxonomies, Xu et al. 2023b generalize them to two layers of task-agnostic categories and introduce task-specific categories for COC rationale annotation. In this work, we study judges’ disagreements in case decision votes, presenting a taxonomy of disagreement with SV-specific subcategories.

**Perceived Difficulty** Regarding model difficulty, there is increasing interest in identifying difficult data instances. Various techniques, such as Influence Functions (Koh and Liang, 2017) and training loss (Han et al., 2018) have been proposed to identify the difficulty of data instances to a certain model. Pointwise V-usable information (PVI) is a recently introduced difficulty metric (Ethayarajh et al., 2022), which incorporates mutual information and other types of informativeness (Xu et al., 2020). Despite its recent introduction, PVI has gained significant attention and proven effective in various tasks, including rationale evaluation (Prasad et al., 2023) and data selection for augmentation (Lin et al., 2023). In this work, we leverage PVI to evaluate the alignment of perceived difficulty between human and COC models.

**Calibration** Recently in mainstream NLP, researchers posit that the overall reliability of a model is determined by two sides: 1) trustworthiness, which is addressed through models’ confidence measuring, and 2) fairness, which is addressed through their confidence alignment with humans (Baan et al., 2024). Existing calibration studies primarily focus on a classifier’s confidence of its predictive performance, commonly evaluated against

the human majority class (Guo et al., 2017; Desai and Durrett, 2020). Baan et al. 2022 instead argue that calibrating against the human majority class is not meaningful in settings with inherent HLV. In our work, we assess the performance and human calibration of COC models in the context of split-vote, a naturally occurring HLV. To the best of our knowledge, this is the first systematic exploration of human calibration in the legal domain.

**Uncertainty Evaluation** Within the deep learning community, uncertainty is often classified into two types: aleatoric and epistemic uncertainty. Epistemic (or model) uncertainty arises from a lack of knowledge about the best model, often exacerbated by out-of-distribution examples. Aleatoric uncertainty, on the other hand, stems from inherent ambiguity and can be considered as the variability in experiment outcomes (Houlsby et al., 2011; Gal et al., 2016). As SV is due to inherent disagreement among judges, we regard it as aleatoric uncertainty. Aleatoric uncertainty is typically quantified through metrics such as Entropy (Gal et al., 2016) and Softmax Response (Geifman and El-Yaniv, 2017), as recognized in prior works (Malinin and Gales, 2018). However, studies show that these methods are based on predictive entropy, and actually measure total uncertainty, combining both epistemic and aleatoric uncertainty (Van Amersfoort et al., 2020). Only when we have prior knowledge that either aleatoric or epistemic uncertainty is low, can we use predictive entropy as a suitable measure for the other type (Mukhoti et al., 2023). Given our small, label-imbalanced, and temporally shifted dataset, we choose to refrain from utilizing these methods to directly measure the aleatoric uncertainty of our cases. Instead, we opt to assess difficulty using PVI, which measures predictive entropy and can be considered an equivalent for total uncertainty.

### 3 Legal and Linguistic Backgrounds

**The Legal Decision Making Process** at the ECtHR begins with the applicants lodging their accusation, alleging one or more violations of articles of the European Convention on Human Rights (ECHR). This process aligns with Task B (allegation prediction) in the widely used LexGLUE benchmark (Chalkidis et al., 2022a). After receiving the complaint, the court undertakes a review of the case, aiming to determine whether a violation has indeed occurred, corresponding to Task A

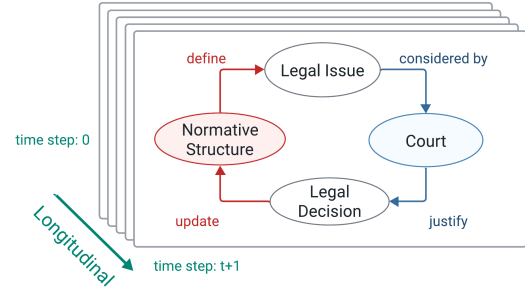


Figure 2: Schematic Diagram of Sources of Disagreement in Legal Decision Process

(violation prediction) in LexGLUE. Cases falling under the purview of well-established ECtHR case-law are directed to a three-judge *Committee*, while others may find themselves before a seven-judge *Chamber*, where decisions are reached through a majority vote. In some circumstances, the *Grand Chamber*, comprising 17 judges, adjudicates cases referred to it by request. When rendering a judgment, the Court typically examines only the specific articles alleged by the applicant. This procedure aligns with Task A1B (violation prediction given allegation; Santosh et al. 2022). Our study evaluates the uncertainty of COC models for Task A1B, mirroring the real legal process.

#### Sources of Disagreement in Legal Decisions

Recent research has delved into sources of HLV and disagreement across various NLP tasks from a linguistic perspective. The classic framework “Triangle of Reference” (Ogden and Richards, 1923) is widely adopted to categorize disagreement sources in classification tasks (Aroyo and Welty, 2015; Jiang and Marneffe, 2022). This concept primarily addresses the relationship between linguistic symbols and the corresponding objects they represent. To study uncertainty sources in NLG tasks, Baan et al. 2023 extend it to the “Double Triangle of Language Production” catering to the complexities of language generation. However, the legal decision-making process introduces an additional layer of complexity. Legal scholars conceptualize the decision-making process in case-law as circular (Ichim, 2019). To better capture this nature, we propose the adoption of the “Direction of Fit” (DoF) framework from speech act theory (Searle and Vanderveken, 1985). In our proposed DoF framework for the legal decision-making process (Fig 2), judges render decisions based on the interpretation of the existing normative structure of case-

law (Law-to-Case). Simultaneously, these case decisions serve as a foundation for future litigation (Case-to-Law). We extend this model with a temporal axis to capture the continuous reproduction of the main normative model over time.

## 4 The SV-ECHR Dataset

**Dataset Collection** We extract the judge votes distribution for each alleged article<sup>\*</sup> from the public database HUDOC<sup>\*</sup> using regular expressions. The information about judge votes is always present in the conclusion section and generally follows certain patterns (See more details in [App A](#)). We did a two-round quality assessment.<sup>\*</sup> The F1 score of our regular expression rules increased from 0.81 in the first round to 0.98 in the second round.

**Dataset Analysis** We then augment the ECHR AIB dataset ([Santosh et al., 2022](#)) with our collected vote information according to the document ID. We name the augmented dataset SV-ECHR. It consists of 11k case fact descriptions along with target label information about which convention articles have been alleged to be violated (task B), and which the court has eventually found to be violated (task A), and the judges vote distribution of each alleged article. The dataset is chronologically split into training (2001–2016), validation (2016–2017) and test set (2017–2019) with 9k, 1k, 1k cases respectively. The label set includes 10 prominent ECHR convention *articles*. On average, each *case* has around 1.6 *alleged articles*. Among all 17,604 *alleged case-alleged article pairs* (hereafter *pair*), only around 7% are split-voted. These numbers underscore the significant label imbalance within the dataset. Additional statistics can be found in [Tab 4](#) in [App C](#). To investigate the extent of judges’ disagreement in SV-ECHR, we assess the entropies of vote distributions for each *pair*. For the detailed calculation and histogram of the entropy distribution, see [App E](#). We see a large share ( $\sim 60\%$  of *chamber votes*) of *single dissenting votes*, where only one single judge voted differently than the six judge majority. A similar pattern is observed for *Grand Chamber* cases involving 17 judges. This pattern aligns with other observations made in empirical legal scholarship ([Fobbe 2022](#), see [App F](#)).

<sup>\*</sup>Following [Chalkidis et al. 2022a](#); [Santosh et al. 2022](#) We use only the 10 most prominent ECHR articles.

<sup>\*</sup><https://hudoc.echr.coe.int>

<sup>\*</sup>[App B](#) offers details on the quality assessment process.

**Correction of Inconsistent Metadata** [Xu et al. 2023b](#) already pointed out inconsistent allegation information in HUDOC. During our quality control process, we found further inconsistent violation information in approximately 2% of the training set. We updated SV-ECHR with the correct metadata.<sup>\*</sup> Our finding of such inconsistencies calls for mindful data curation when developing COC datasets.

## 5 Disagreement among Judges’ Votes

### 5.1 Disagreement Taxonomy

We use the disagreement taxonomy from [Xu et al. 2023b](#) to analyze the reasons behind judges’ split votes. [Fig 3](#) displays the taxonomy, with two adapted task-agnostic levels and our expanded split-vote specific subcategories and proxy variables.<sup>\*</sup> In the following sections, we provide detailed explanations for each SV-specific category.

#### 5.1.1 The Data

Disagreements among judges can stem from various case aspects, including *Genuine Ambiguity* within the Normative Structure, *Narrative Complexity* of the facts, and/or *Specific Legal Context*.

**Genuine Ambiguity** is attributed to **Normative Uncertainty** of case law in the context of legal NLP ([Xu et al. 2023b](#)), which emerges when the court is presented with the possibility of justifying an outcome through multiple legal source interpretations and argumentation. Its occurrence is not uncommon in ECtHR judgments due to the deliberate drafting of the convention in a ‘flexible’ manner to allow tailoring the interpretation to domestic specificities based on the subsidiarity and margin of appreciation principles. It should be noted that there exists an analog *factual* uncertainty where the facts of the case are unclear based on limited (or contradicting) evidence provided. As the ECtHR does not engage in evidentiary reasoning, this aspect is out of scope for this work.

**Text Instantiation** covers inconsistency, incompleteness, or biases during text production (i.e. judgment document drafting). When reviewing case files, judges analyze materials that encapsulate the factual background, legal arguments, and evidence presented by both parties. However, these

<sup>\*</sup>See [App D](#) for more details of our metadata correction.

<sup>\*</sup>For a comprehensive understanding of each taxonomy category, we direct the reader to [Xu et al. 2023b](#)



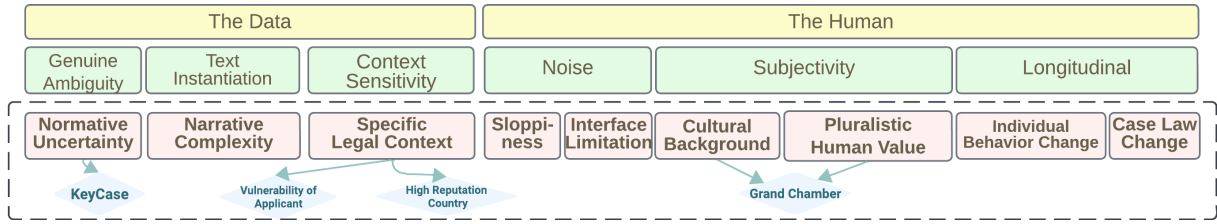


Figure 3: Taxonomy of disagreement sources among judges, with two adapted task-agnostic levels from Xu et al. 2023b and our expanded split-vote specific subcategories and proxy variables (in dashed box).

documents are usually produced by the Court Registry, priming the language accordingly. The likelihood of encountering textual unclarity, such as inconsistent framing of facts, increases with the length and complexity of the docket. We refer to this challenge as **Narrative Complexity**.

**Context Sensitivity** refers to the **Specific Legal Context**, characterized by undefined and controversial social-political factors for which judges may not find a precisely tailored legal explanation. We use the term ‘Context Specificity’ because pinpointing the exact arrangement of factors (i.e., attitudinal, normative, and strategic) influencing context in a specific decision is challenging (Shapiro, 2002). Certain social and political factors may only become contentious in a particular context and arrangement of factors.

### 5.1.2 The Annotator/Judge

Disagreements can also arise due to variation in human behavior, which we systematize as *Noise*, *Subjectivity* and *Longitudinal* behavioural change. *Noise* covers errors due to annotator’s *Sloppy Annotation* or *Interface Limitation* (Uma et al., 2021; Sandri et al., 2023). We expect negligence-related noise and interface limitations to be insignificant in naturally occurring ECtHR judge votes.\*

**Subjectivity** The ECtHR is composed of judges from the 46 member states with diverse legal traditions and cultural backgrounds. Previous work in general NLP demonstrated how annotators’ demographic identities can influence how they label toxicity in text (Sap et al., 2022). The judges’ **Cultural Background**, stemming from their legal training and political consideration from their native countries may lead to variations in their rulings com-

pared to those of their colleagues (Voeten, 2008).<sup>\*</sup> Divergent opinions may arise when, for example, two judges from the same country prioritizing different values. **Pluralistic Human Values** have recently been made a primary object of mainstream NLP research (Sorensen et al., 2024). Political science research has explored this in the context of quantitative justice ideal point estimation on the US Supreme Court (SCOTUS, typically a scalar dimension between liberal and conservative, see Segal and Cover 1989, Martin and Quinn 2002). More similar to our focus on agreement, Ruger et al. 2004 trained an ensemble of court-level and judge-specific decision trees, and observed different performance on conservative and liberal judges.

A comparable liberal vs conservative investigation in the ECtHR context would not be useful. It is composed of 46 Member States, each with its own spectrum of political ideologies, more pluralist and complicated than the liberal-conservative divide. Moreover, there is no temporal stability in the judges’ behavior because they have a limited mandate and act in different bench formations (chamber, grand chamber, etc.). Focusing on individual ideologies to the exclusion of variables accounting for the wider court dynamics could not support accurate analysis.

**The Longitudinal Dimension** in Xu et al. 2023b considers two aspects: **Individual Behavioral Change** over time and **Case Law Change**. The latter involves shifts in the collective societal attitude towards specific phenomena over time, including changes in laws, policies, and cultural norms. In the legal domain, legal professionals must continually adapt their knowledge, strategies and reasoning to align with the ongoing evolvement of jurisprudence and applicable statutory law. Due to the limited scope of this work, we defer a comprehensive longitudinal study to future research.

\*Empirical legal scholarship has been investigating the relationship between judge productivity and caseload, including its ramifications on decision quality (e.g., Engel and Weinshall 2020). To the best of our knowledge, no such evidence has been documented in the ECtHR context.

\*See discussion about *National Judge* in App H

## 5.2 Conflation of Categories

It is important to acknowledge that our taxonomy and its categories are not always sharply delineated. Conflation can occur and the boundary between categories may be blurry. Some sources can be interpreted as either Law-to-Case or Case-to-Law. For instance, cases involving vulnerable applicants often involve complex social-political issues lacking clear legal precedent, causing disagreement on the bench. Simultaneously, the judges’ subjectivity may lead to varying opinions on whether an applicant should be deemed vulnerable, i.e. judge subjectivity and legal context interact.

## 5.3 Proxies & Disagreement Correlation

To quantitatively explore the influence of different taxonomic categories on judges’ votes, we work with a legal expert to identify proxy variables. We hypothesize that these proxies correspond to higher disagreement among judges and evaluate the statistical association between them and the entropy of the vote distribution. If not otherwise mentioned, we retrieve relevant information from the HUDOC dataset and label a case as 1 if it is listed with the proxy feature and 0 otherwise.

For **Normative Uncertainty**, we choose *Key-Case* as a proxy following Xu et al. 2023b. The court annually chooses a set of “key cases”, which often deal with complex and novel legal issues. Given the absence of established legal standards for them, they often generate controversy and disagreement among experts. We hypothesize that judges tend to disagree more in “key cases”.

**Specific Legal Context** via two proxy variables: *HighRepCountry* (High Reputation Country): Previous legal and political science work indicates ECtHR judges tend to have more split votes in cases where the defending country is regarded as a High Reputation Country\* (Dothan, 2014).

*VulnApplicant* (Vulnerability of Applicant): the ECtHR court adapts convention standards to meet vulnerable individuals’ needs despite *vulnerability* not being defined in the Convention. It remains underdefined within the context of ECtHR because the judges do not explain the process through which they identify an individual as vulnerable (Heri, 2021), which can lead to divergent legal opinions. We retrieve relevant information from the VECHR dataset (Xu et al., 2023a) and categorize a case as

1 if its applicant is regarded as vulnerable.\*

**Subjectivity** refers to the disagreement due to personal opinions and values. We propose that the 17-judge *Grand Chamber* is more prone to disagreement than the 7-judge *Chamber*. This is due to the inherent pluralism of human values in larger groups, which naturally fosters a broader range of viewpoints and thus, increased subjectivity. As explained in sec. 5.1.2, modeling specific ideological dimensions of individual judges is out of scope for this work. We account for some limited political dynamics by including the democratic score of the state against which the claim is brought (the *HighRep* variable above).

### 5.3.1 Do proxy measures correlate with judges’ votes?

To measure the influence of each selected taxonomy category, for each binary proxy variable, we compute the entropy of the vote distribution among all cases exhibiting that variable (‘present’, value 1) and those that do not (‘absent’, value 0). We perform an independent t-test to compare the mean entropies between the two groups.

	mean entropy		t-value	p-value
Proxy (absent/present):	0	1		
GrandChamber	0.50	0.49	1.61	0.11
HighRepCountry	0.48	0.51	-4.28	2e-5*
VulnApplicant	0.50	0.62	-1.8871	0.006*
KeyCase	0.50	0.48	2.46	0.002*

Table 1: Associations between proxy variables and vote distribution entropy. \*:  $p < 0.05$ . Green highlighted: confirming our hypothesis. Red highlighted: contradicting our hypothesis.

Tab 1 shows significantly lower agreement among judges’ votes for *HighRepCountry* and *VulnApplicant* ( $p$ -value  $< 0.05$ ). In other words, specific contexts of case facts correlate more with disagreement than Normative Uncertainty. Judges express disagreement related to how facts fit the norm because of circumstances related to the social standing of the applicant and the political situation of the defending State. This is an inherent characteristic of adjudication where judges qualify the context so as to fit the norm. The inclusion of a category related to the legal norm’s scope could have provided further insights.

\*Only annotation of cases under Article 3 are available in VECHR before 2016. Hence our t-test of *VulnApplicant* is only on cases of Article 3. App K offers more explanation on Vulnerability in ECtHR jurisprudence.

\*Countries with higher democracy index. See App J.

Our hypothesis regarding *GrandChamber* does not hold precisely because it does not matter which bench formation deals with the case - Chamber or Grand Chamber. What matters is how judges decide to fit the facts to the norm.

Interestingly, we found less disagreement among judges on *KeyCase*. Experts suggest this may stem from the Court’s assigning the importance score ex post, after interpretation and agreement are reached. The Court is more likely to designate a consensus case as *KeyCase* to maintain coherent jurisprudence, rather than a controversial case that could invite future applicants to base cases on disputed facts and dissenting opinions. As reflected in our DoF framework (??), Case-to-Law uncertainty inherently perpetuates Law-to-Case uncertainty.

## 6 Judge-Model Misalignment

### 6.1 Experimental Setup of COC

We assess the alignment of a model’s prediction difficulty\* with judge disagreement, specifically COC models trained on ECtHR Task A1B. To address input length constraints, we employ a LegalBERT (Chalkidis et al., 2020) variant of the hierarchical attention model\* (Yang et al., 2016), as adapted from Santosh et al. 2022.

**Task** Benchmark task ECtHR A1B (Violation Identification given Allegation) identifies the set of violated convention article(s) from the textual case facts and list of allegedly violated articles.

**Models** Santosh et al. 2022 encodes allegation information as a multi-hot vector concatenated with the text representation. By contrast, in this study, we textify allegation information as “Alleged Article: X”, where X is a comma-separated list of articles and prepend to the input text. We employ the LegalBERT (nlpaueb/legal-bert-base-uncased; Chalkidis et al. 2020) as the backbone in hierarchical models.\*

**Metrics** Our evaluation of COC performance employs the micro-F1 (mic-F1), macro-F1 (mac-F1), and hard-macro-F1 (hm-F1; Santosh et al. 2022). The hm-F1 is calculated as the mean F1-score for

\*Another approach would be assessing a model’s prediction *uncertainty*. Refer to § 2 for an explanation of our decision to abstain from using Uncertainty evaluation methods

\*See App N for model structure and implementation.

\*To explore the impact of domain-specific pre-training, we did also experiment on BERT. See App L.

each article, where cases with that article having been violated are considered as positive instances, and cases with that article being alleged but not found to be violated as negative instances, resulting in a smaller pool of more difficult negatives.

**COC Performance** We run experiments across three random seeds. The hierarchical LegalBERT achieves  $73.55 \pm 0.62$  for mic-F1,  $68.75 \pm 1.13$  for mac-F1 and  $68.03 \pm 1.29$  for hm-F1. Although the mic-F1 and mac-F1 scores closely align with those reported in (Santosh et al., 2022; Xu et al., 2023b), differences of  $\sim 2\%$ , the hm-F1 performance demonstrate a noteworthy improvement of  $\sim 10\%$ .\* Importantly, hm-F1 exhibits particular sensitivity to allegation information. For the rest of this paper, we adopt hm-F1 as our primary metric for COC performance.

### 6.2 Difficulty of SV Cases in COC

Pointwise V-usable information (PVI; Ethayarajh et al., 2022) measures the difficulty of an instance within a dataset for a given model as:  $PVI(x \rightarrow y) = -\log_2 g[x](y) + \log_2 g'[\emptyset](y)$ . Computing PVI involves fine-tuning a model  $g$  on two datasets. The first dataset  $D$  comprises input-target pairs  $\{(x_i, y_i) | (x_i, y_i) \in D\}$ , while the second  $D'$  is used to fine-tune model  $g'$  on null-target pairs,  $\{(\emptyset, y_i) | (x_i, y_i) \in D\}$  ( $\emptyset$  represents an empty string intended to fit the label distribution). PVI serves as the measure of information gain resulting from the provision of an input during fine-tuning. Higher PVI suggests a better representation of the input in the model, and thus an easier instance.

**Experiment Settings** We use the COC fine-tuned models from Sec 6.1 as the input-target model  $g$ . We fine-tune an architecturally identical model  $g'$  with the input replaced by  $\emptyset$ . Subsequently, we partition the test set into two distinct partitions of SV and unanimous cases, respectively. Our expectation is that the SV subset will exhibit lower average PVI compared to the unanimous subset. We use entropy of the judges’ vote distribution as an estimator of case difficulty: the more the judges disagree with each other, the harder a case is.

**Results** Tab 3 presents the average COC performance (hm-F1) and model-perceived difficulty (mean PVI scores). The model exhibits lower hm-

\*This can be partly attributed to our correction of labels (see § 4). Some variation may also be due to randomness in weight initialization. See our discussion in App M

		hm-F1 $\uparrow$			ECE $\downarrow$			DistCE $\downarrow$			count
		/	TS	soft	/	TS	soft	/	TS	soft	
LegalBERT	u	<b>69.30 <math>\pm</math> 1.88</b>	<b>69.30 <math>\pm</math> 1.88</b>	67.31 $\pm$ 0.5	23.32 $\pm$ 1.01	<b>2.95 <math>\pm</math> 0.67</b>	22.04 $\pm$ 0.63	25.10 $\pm$ 1.17	37.23 $\pm$ 1.71	<b>24.70 <math>\pm</math> 0.17</b>	1463
	sv	<b>53.67 <math>\pm</math> 4.59</b>	<b>53.67 <math>\pm</math> 4.59</b>	46.21 $\pm$ 4.41	29.92 $\pm$ 2.81	<b>8.02 <math>\pm</math> 0.58</b>	28.49 $\pm$ 0.82	41.03 $\pm$ 2.75	<b>28.28 <math>\pm</math> 0.99</b>	40.72 $\pm$ 1.00	112
	all	<b>68.03 <math>\pm</math> 1.29</b>	<b>68.03 <math>\pm</math> 1.29</b>	65.16 $\pm$ 0.82	23.75 $\pm$ 1.08	<b>2.99 <math>\pm</math> 0.83</b>	22.32 $\pm$ 0.64	26.23 $\pm$ 1.0	36.60 $\pm$ 1.52	<b>25.84 <math>\pm</math> 0.21</b>	1575

Table 2: COC performance (hm-F1), confidence-calibration (ECE), and human-calibration (DistCE) performance with std ( $\pm$ ) on test set. "/": COC finetuned models in § 6.1; "soft": models fine-tuned with soft-loss; "TS": models after Temperature Scaling. Results shown over 3 random seeds. See Tab 8 in App Q for performance on dev set.

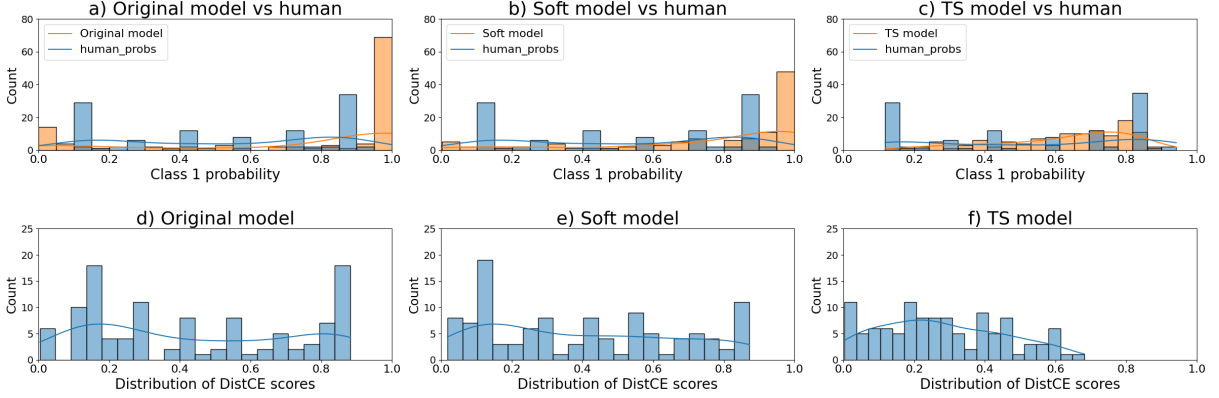


Figure 4: The distributions over probabilities for class 1 of the models vs human vote distributions (row 1) and distCE (row 2). See Fig 8 in App Q for more figures comparing human uncertainty to model uncertainty

	hm-F1		mean PVI		t-value	p-value
	U.	SV	U.	SV		
LegalBERT	69.3 $\pm$ 1.88	53.67 $\pm$ 4.59	0.77	-0.12	2.66	0.008

Table 3: COC prediction (hm-F1) and difficulty scores (mean PVI) for unanimous (U.) and SV cases.

F1 and PVI scores for SV cases compared to unanimous cases. An independent t-test on the average PVI scores between the two groups reveals a significant difference ( $p < 0.05$ ). This indicates that SV cases are more challenging for the models than unanimous cases, aligning with human perceptions of difficulty. We also calculate the Pearson correlation coefficients between PVI scores and the entropy of judges’ vote distribution of SV cases. LegalBERT has a correlation coefficient (r-value) of -0.068 and a p-value of 0.48. The results reveal a negative correlation between PVI scores and the degree of disagreement among judges, consistent with our expectations. However, the observed correlation is very weak ( $|r| < 0.1$ ) and statistically insignificant ( $p > 0.05$ ). This suggests that, while the model can capture differences in difficulty between unanimous and SV cases, it struggle to accurately represent the nuanced degree of judges’ disagreement. Therefore, we next evaluate the models’ confidence and soft-label training.

### 6.3 Calibration of SV Cases

#### 6.3.1 Evaluation Metrics

Most current NLP research focuses on **Confidence Calibration** (Jiang et al., 2021; Desai and Durrett, 2020): *The model should be unsure when it does not know the answer.* A model is considered well confidence-calibrated if its prediction confidence aligns with its predictive accuracy, commonly evaluated against the human majority class. Yet there is a growing interest in accounting for HLV and/or pluralistic values (Plank, 2022; Sorensen et al., 2023). Therefore we extend our evaluation to include **Human Calibration** (Baan et al., 2022): *The model should be unsure when humans are unsure about the answer.* We consider a model well human-calibrated if the categorical distribution of predicted class probabilities align well with the actual human vote distribution.

#### 6.3.2 Calibration Methods

For *confidence calibration*, we employ **Temperature Scaling** (TS; Guo et al. 2017). This simple yet widely used post-hoc method uses a single temperature parameter  $t$  to scale the output logits of a classifier. We choose the temperature  $t$  by searching a range of possible values for  $t$  on the dev set. For *human calibration*, we adopt the approach from Peterson et al. 2019; Uma et al. 2020 by finetun-



ing with a **Soft Loss Function**.<sup>\*</sup> During training, models are exposed to “soft labels” derived from the judges’ vote distributions, serving as target distributions in a cross-entropy loss function.

**Expected Calibration Error** (ECE) is the most often used metric for *confidence calibration* (Naeini et al., 2015; Guo et al., 2017). A lower ECE indicates better calibration, suggesting that the model’s predicted probabilities are more accurate reflections of the true probabilities. Refer to the App P for further details on ECE.

**DistCE** by Baan et al. 2022 measures *human calibration*, which can be calculated as  $\text{DistCE}(x) = 1/2\|q - p\|$ , where  $q$  is the vote distribution and  $p$  is the model’s predictive distribution.

### 6.3.3 Results & Discussion

Tab 2 presents the experimental results. We make the following observations:

**i) SV cases are indeed the most challenging.** It is reflected by their lowest COC performance (hm-F1). They also exhibit a higher degree of miscalibration (higher ECE) and misalignment with human responses (higher distCE).

**ii) Applying TS greatly enhanced confidence calibration.** Tab 2 presents that applying TS has greatly reduced the ECE score, without negatively impacting COC performance (hm-F1).

**iii) The models remain misaligned with the human vote distribution.** Tab 2 shows that soft-loss tuning only slightly improves the alignment between the model’s prediction confidence and the human vote distribution (lower DistCE). Fig 4 illustrates that Soft LegalBERT exhibits noticeably fewer instances of over-confident predictions when judges do not unanimously agree, as depicted on near-0/1 probability portions of Fig 4b, in contrast to the original model shown in Fig 4a. On split-vote cases, TS models exhibit substantially lower DistCE scores than soft models. However, we do not consider that TS provides a better human calibration. The lower DistCE scores may be attributed to an overly aggressive temperature, resulting in a more uniform output distribution (see Fig 4c, with a temperature of 5.5). Moreover, judge votes often exhibit a quasi-bimodal distribution, with many split votes caused by a single judge dissenting pro/con the finding of a violation (see discussion

on *single dissenting votes* in § 4). Further, the DistCE score histograms (Fig 4d-f) illustrate that soft-tuning (Fig 4e) improves human alignment by reducing instances of extreme miscalibrations in the right tail, as compared to the original model in Fig 4d. It is noteworthy that the TS model (Fig 4f) shifts the distribution further to the left by reducing the right tail. However, this is at the expense of predictions that perfectly align with human judgement probabilities, as evidenced by fewer instances of DistCE scores below 0.15 in the TS model compared to the Soft model. Therefore, despite TS models potentially displaying lower DistCE scores than soft models, they do not provide an optimal fit. It is crucial to analyze the distributions prior to drawing conclusions, especially as we have shown in cases with split votes, where distributions are *bimodal*.

## 7 Conclusion

We present SV-ECHR, a new COC dataset enriched with naturally occurring split-votes by judges on the ECtHR. We also present a SV-enriched taxonomy of disagreement sources. Our experiments reveal shortcomings when combining TS and ECE to improve and measure calibration of models against labels subject to inherent human disagreement. This is due to a distribution mismatch reflected in low DistCE scores. Soft loss training produces only slightly better human calibration scores. We call on the community to explore methods for measuring and improving the alignment of model calibration with human behavior; as well as more research into incorporating HLV in NLP.

## Limitations

Our study is constrained by the datasets, models, and selective prediction techniques under consideration, primarily relying on the ECtHR dataset. Expanding the investigation to encompass diverse datasets and legal jurisdictions would enhance our understanding of disagreement in judge decision votes and the alignment of perceived difficulty between judges and models.

Additionally, due to computational limitations, we are constrained from pre-training language models from scratch or fine-tuning large language models (LLMs). Our study relies on existing pre-trained BERT-based models, focusing solely on fine-tuning. We refrain from exploring LLM models, as no widely agreed-upon method for measur-

<sup>\*</sup>See App O for details of the Soft Loss Function.

ing calibration for LLMs has emerged at the time of submission. Furthermore, with respect to variations introduced by prompts and data contamination during pretraining, exploring the use of LLMs for difficulty perception and calibration on a small-scale, specialized legal dataset is a distinct research question deserving a separate paper.

## Ethics Statement

In this study, our retrieved judges' vote information from the publicly available HUDOC dataset, with the overarching goal of improving the alignment of perceived difficulty and calibration between models and judges. While these decisions include real names and are not anonymized, we do not anticipate any harm beyond the availability of this information resulting from our experiments.

The task of case outcome classification raises significant ethical and legal concerns, both generally and specifically concerning the European Court of Human Rights (Medvedeva et al., 2020). It is important to clarify that we do not advocate for the practical implementation of COC within courts. Previous work (Santosh et al., 2022) has demonstrated that these systems heavily rely on superficial and statistically predictive signals lacking legal relevance. This underscores the potential risks associated with employing predictive systems in critical domains like law and highlights the importance of trustworthy and explainable legal NLP.

In this work, we investigate the sources of disagreements among judges in their decisions on case facts. While technically situated outcome classification models, we intend our analysis of different types of disagreement to promote the acceptance of human label variation and pluralistic human values within the legal NLP community. By acknowledging and understanding various perspectives, interpretations, and biases of judges, we contribute to a more comprehensive and inclusive discourse within the field. A continuation of this work can unfold long term practical implications: By identifying patterns which create uncertainty, applicants could potentially 'exploit' the distinctive circumstances that change the normative assessment for the purpose of their own case, or otherwise inform litigation strategy. By the same token, however, it would offer the court the possibility to investigate and 'check' whether it applies legal norms coherently, in line with the demands of consistency, foreseeability and certainty. The potential effects

of data-driven decision making in the legal domain cut both ways, and must be reconciled mindfully.

## Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and suggestions. We would also like to thank the members of the MaiNLP/CIS research lab for their thoughtful feedback. We acknowledge the support of our funders. BP is supported by the ERC Consolidator Grant No. 101043235.

## References

- Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel PreoŃiuc-Pietro, and Vasileios Lamps. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.
- Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.
- Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernández. 2022. [Stop measuring calibration when humans disagree](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.
- Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. [Interpreting predictive probabilities: Model confidence or human label variation?](#) In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277, St. Julian's, Malta. Association for Computational Linguistics.
- Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*.
- Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of ACL 2019*, 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 EMNLP 2020*, pages 2898–2904.
- 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 ACL 2022*, pages 4310–4330.
- Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022b. [FairLex: A multilingual benchmark for evaluating fairness in legal text processing](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.
- Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.
- Shrey Desai and Greg Durrett. 2020. [Calibration of pre-trained transformers](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.
- Shai Dothan. 2014. *Reputation and Judicial Tactics. A Theory of National and International Courts*. Oxford University Press.
- Ran El-Yaniv et al. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5).
- Christoph Engel and Keren Weinshall. 2020. Manna from heaven for judges: Judges’ reaction to a quasi-random reduction in caseload. *Journal of Empirical Legal Studies*, 17(4):722–751.
- Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. [Understanding dataset difficulty with  \$\mathcal{V}\$ -usable information](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.
- Sean Fobbe. 2022. Introducing twin corpora of decisions for the international court of justice (icj) and the permanent court of international justice (pcij). *Journal of Empirical Legal Studies*, 19(2):491–524.
- Yarin Gal et al. 2016. *Uncertainty in deep learning*. PhD thesis, University of Cambridge.
- Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.
- 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.
- Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. 2018. Co-teaching: robust training of deep neural networks with extremely noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, page 8536–8546, Red Hook, NY, USA. Curran Associates Inc.
- Laurence R Helfer and Erik Voeten. 2020. Walking back human rights in europe? *European Journal of International Law*, 31(3):797–827.
- Corina Heri. 2021. *Responsive Human Rights: Vulnerability, Ill-treatment and the ECtHR*. Bloomsbury Academic.
- Daniel E Ho, Cassandra Handan-Nader, David Ames, and David Marcus. 2019. Quality review of mass adjudication: A randomized natural experiment at the board of veterans appeals, 2003–16. *The Journal of Law, Economics, and Organization*, 35(2):239–288.
- Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.
- Oana Ichim. 2019. *The European Court of Human Rights between Dispute Settlement and Governance*. PhD thesis, Graduate Institute of International and Development Studies.
- Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [How can we know when language models know? on the calibration of language models for question answering](#). *Transactions of the Association for Computational Linguistics*, 9:962–977.
- 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*, pages 4171–4186.
- Pang Wei Koh and Percy Liang. 2017. [Understanding black-box predictions via influence functions](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.
- Martin Kuijer. 1997. [Voting behaviour and national bias in the european court of human rights and the international court of justice](#). *Leiden Journal of International Law*, 10(1):49–67.



- Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazi-far, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023. [Selective in-context data augmentation for intent detection using pointwise V-information](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1463–1476, Dubrovnik, Croatia. Association for Computational Linguistics.
- Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31.
- Andrew D Martin and Kevin M Quinn. 2002. Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999. *Political analysis*, 10(2):134–153.
- Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28(2):237–266.
- Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. 2023. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394.
- Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.
- Charles Kay Ogden and Ivor Armstrong Richards. 1923. *The meaning of meaning A study of the influence of thought and of the science of symbolism*. Harcourt, Brace & World, Inc.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Ellie Pavlick and Tom Kwiatkowski. 2019. [Inherent disagreements in human textual inferences](#). *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. [ReCEval: Evaluating reasoning chains via correctness and informativeness](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086, Singapore. Association for Computational Linguistics.
- Theodore W Ruger, Pauline T Kim, Andrew D Martin, and Kevin M Quinn. 2004. The supreme court forecasting project: Legal and political science approaches to predicting supreme court decisionmaking. *Columbia law review*, pages 1150–1210.
- Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. [Why don’t you do it right? analysing annotators’ disagreement in subjective tasks](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.
- T.y.s.s 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 EMNLP 2022*.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.
- John R Searle and Daniel Vanderveken. 1985. *Foundations of illocutionary logic*. CUP Archive.
- Jeffrey A Segal and Albert D Cover. 1989. Ideological values and the votes of us supreme court justices. *American Political Science Review*, 83(2):557–565.
- Martin Shapiro. 2002. [19Political Jurisprudence](#). In *On Law, Politics, and Judicialization*. Oxford University Press.
- Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2023. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. *arXiv preprint arXiv:2309.00779*.
- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.



- Zeeraq Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
- Santosh T.y.s.s, 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 605–617, Dubrovnik, Croatia. Association for Computational Linguistics.
- Santosh T.y.s.s, 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.
- Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.
- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. [Uncertainty estimation using a single deep deterministic neural network](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9690–9700. PMLR.
- Erik Voeten. 2008. The impartiality of international judges: Evidence from the european court of human rights. *American Political Science Review*, 102(4):417–433.
- Keren Weinsahl-Margel and John Shapard. 2011. Overlooked factors in the analysis of parole decisions. *Proceedings of the National Academy of Sciences*, 108(42):E833–E833.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Trans-formers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Shanshan Xu, Leon Staufer, Santosh T.y.s.s, Oana Ichim, Corina Heri, and Matthias Grabmair. 2023a. [VECHR: A dataset for explainable and robust classification of vulnerability type in the European court of human rights](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11738–11752, Singapore. Association for Computational Linguistics.
- Shanshan Xu, Santosh T.y.s.s, Oana Ichim, Isabella Risini, Barbara Plank, and Matthias Grabmair. 2023b. [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.
- Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*, pages 1480–1489.

## A Information about Judge’ Votes in HUDOC

We observe that the information about judge’ votes is always present in conclusion section of the decision towards the end and generally following certain patterns, such as: "Holds, by X votes to Y, that there has been a/no violation of Article Z of the Convention"; or "Holds by X votes to Y that Article Z has (not) been violated" as in [Fig 5](#).

## B Quality Validation of SV-ECHR

To ensure the quality of the dataset, we did a two-round quality assessment. In the first round, we manually retrieved the split-vote information of 20 cases as gold labels. We then assess the votes information exacted by our regex rule with the gold labels. We analysis the error and improve our regex rule. In the second round, we repeat this process with another 20 cases. The F1 score increased from 0.81 in the first round to 0.98 in the second round.

## C Dataset Statistics

[Tab 4](#) offers additional statistics on the SV-ECHR dataset.

## FOR THESE REASONS, THE COURT

1. *Declares*, by a majority, the complaint under Article 6 of the Convention admissible and the remainder of the application inadmissible;
2. *Holds*, by nine votes to eight, that there has been no violation of Article 6 § 1 of the Convention.

Figure 5: Screenshot of Judges’ vote information in HUDOC

	# Cases	# Case -Article Pairs	# alleged Case- Article Pairs	# SV Case- Article Pairs
Train	9000	90000	14513	960
Dev	1000	10000	1516	135
Test	1000	10000	1575	112

Table 4: More statistics of the SV-ECHR dataset.

	# Case-Article Pairs	# alleged pairs	# pair w/ wrong alleg.	# pair w/ wrong vio.
Train	90000	14513	2719	292
Dev	10000	1516	247	0
Test	10000	1575	278	0

Table 5: Statistics about the corrected meta information.

## D Correction of Inconsistent Metadata

Tab 5 shows the statistics about our correction of case metadata. Notably, apart from the inconsistent allegation information pointed out by Xu et al., 2023b. We further found approximately 2% inconsistent violation information in the train set during our quality control process. Most of the cases are so-called ‘striked-out’ cases which are difficult to parse. We also find finetuning on the corrected train set improves model’s COC performance (§ 6.1) as discussed in App M.

## E Entropy of Judges’ Vote Distribution

Fig 6 display the histogram of the entropy of each pair’s vote distribution. The entropy is calculated as  $H(\mathbf{p}) = -\sum_{i \in \mathcal{C}} p_i \log(p_i)$  and  $p_i = \frac{n_i}{\sum_{j \in \mathcal{C}} n_j}$ , where  $\mathcal{C}$  is the label category set [violation, non-violation] and  $n_i$  is the number of judges voting for category  $i$ .

## F Debates in Legal Scholarship about Single Judge Dissenting Vote

Tab 4 reveals that  $\sim 60\%$  of *chamber* votes (involving 7 judges) exhibit entropy of around 0.4, indicating that only one judge vote differently from the remaining 6 judges (entropy([1,6])  $\approx 0.4$ ). A similar pattern is observed for *Grand Chamber* cases involving 17 judges. This *Single Judge Dissenting* pattern aligns with ongoing debates in legal scholarship. For instance, Fobbe 2022 examined the number of dissenting opinions in decisions from

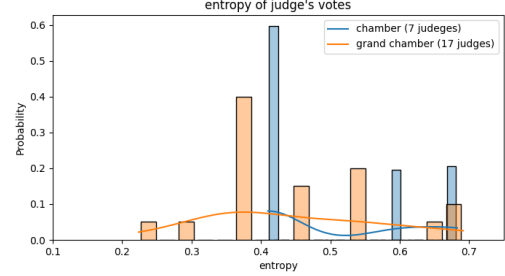


Figure 6: Histogram of entropy of judges’ votes distribution over train/dev/test split

the International Court of Justice (ICJ). Their results indicate a significant proportion of unanimous decisions, followed by a monotonously decreasing number of dissents. Some legal scholars support the view that diverging opinions do not signal a division of the bench inside the Court over the scope of protection of rights, even less a lowering of standards of protection. Others support the contrary view according to which dissenting opinions send a signal of walking-back in terms of effective protection (Helfer and Voeten, 2020). Tab 4 and Fobbe 2022’s finding actually shows that the case law does not consistently, but rather seldomly, give rise to dissenting opinions. Moreover, one can not ignore the national judge’s dissenting vote (App H). It is difficult to assess the weight of a single judge dissenting on the decision of the court as a whole. Often times, dissenting opinions are not followed up on. Our focus here is limited to developing a yardstick against disagreement in human decisions can be measured.

## G Judge’s Negligence Errors

Empirical legal scholarship identified time constraints leading to judge negligence errors in certain jurisdictions like the US Board of Veterans’ Appeals (Ho et al., 2019) and Israeli Magistrate Courts (Engel and Weinshall, 2020). However, to the best of our knowledge, there is no evidence has been documented in the ECtHR context for careless behavior by judges. The infamous Hungry Judge Effect, stemming from work in (Danziger

et al., 2011), suggested that judges tend to issue harsher sentences just before lunch, presumably influenced by hunger. However, a subsequent study (Weinshall-Margel and Shapard, 2011) argues that the observed peak in favorable decisions after a meal break is likely an artifact of case presentation order, considering anticipated outcomes and duration. Overall, this rebuts the notion that hunger impacts judges’ rational decision-making.

## H National Judge

In the ECtHR context, a *national judge* refers to the judge appointed from the respondent country of the case which was brought before the court. Like all other judges, national judges maintain independence and do not act as representatives of their respective governments. However, as previously noted in studies by legal scholars, national judges have been observed to dissent more frequently in cases finding a violation of the Convention compared to their non-national counterparts (Kuijer, 1997; Helfer and Voeten, 2020). Voeten 2008 also provides evidence and explanations, including for why the judges are considered policy-seekers, and concludes that judicial activism is driven by the political logic of European integration.

## I Key Case

*KeyCase*: The ECtHR annually chooses a set of significant cases, known as “key cases”. These often deal with complex and novel legal issues. Given the absence of established legal standards for interpreting them, they often generate controversy and disagreement among judges.

## J High Reputation Country

Previous legal and political science work indicates ECtHR judges tend to have more split vote in cases where the defending country is regarded as a ‘High Reputation Country’. Strong democracies enjoy a high reputation in front of the judges while weak and new democracies only benefit from a low reputation implying that the ECtHR should issue more demanding judgments against low reputation states than against high reputation states, with the view not to damage its own reputation and ensure compliance with its decisions (Dothan, 2014). We extracted the country information from the HUDOC metadata, namely Respondent State(s). Following Chalkidis et al. 2022b, we group the countries according to the disproportion of violations between

eastern and central European countries, and the rest of European countries (western European, Nordic, mediterranean states).

## K Vulnerability in ECtHR

The ECtHR adapts convention standards to meet individual needs and to ensure effective human rights protection. Recognizing vulnerability is crucial for understanding unique needs and implementing targeted support systems. However, the concept of ‘vulnerability’ remains undefined by the court, which can lead to divergent opinions on what qualifies as vulnerability from a legal point of view. Cases involving vulnerable applicants often deal with complex social-political issues related to the protection and rights of individuals who may face various challenges, such as victimization, migration, discrimination, reproductive health, unpopular views etc. The inclusion of vulnerability as a proxy variable is an example of how a human rights legal concept that is difficult to define can potentially give rise to diverging opinions on the bench. A comprehensive study of such concepts in ECtHR jurisprudence lies beyond the scope of this work. We direct the reader to Heri 2021 for a systematic legal study and Xu et al. 2023a for comprehensive NLP research on vulnerability in the ECtHR context.

## L The impact of domain-specific pre-training on uncertainty representations

Tab 6 reports the results of classification performance. Notably, the model with legal-specific pre-training (LegalBERT) outperforms the one with general pre-training (BERT).

Tab 8 shows that fine-tuning with soft-loss to human labels yields minimal ECE changes with a discrepancy: BERT shows a slight decrease, while LegalBERT displays a minor increase. This mirrors the issue identified by Baan et al. 2022, highlighting the challenge of using ECE to measure calibration on disagreement data. They argue that even a classifier perfectly modeling human judgment distribution would still be severely miscalibrated when measured by ECE.

	mic-F1	mac-F1	hm-F1
BERT	72.31 ± 4.09	65.53 ± 6.22	64.22 ± 3.46
LegalBERT	73.55 ± 0.62	68.75 ± 1.13	68.03 ± 1.29

Table 6: COC performance on test set

finetune dataset	mic-F1	mac-F1	hm-F1
ours	<b>73.55 ± 0.62</b>	<b>68.75 ± 1.13</b>	<b>68.03 ± 1.29</b>
Santosh et al. 2022	73.41 ± 2.5	67.74 ± 3.2	63.93 ± 1.7

Table 7: LegalBERT’s COC performance on test set with different finetune dataset. *Our* dataset with correction of metadata as mentioned in § 4

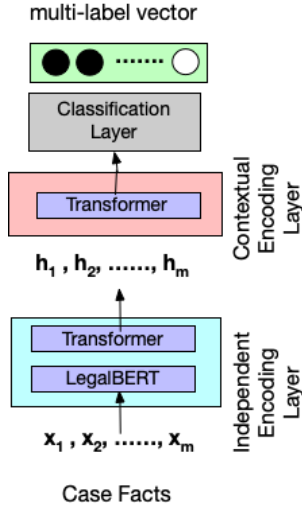


Figure 7: Hierarchical Classification Variant

## M Ablation Experiment

We also train hierarchical LegalBERTs with the textified allegation information on the original dataset from (Santosh et al., 2022) without corrected metadata. Results in Tab 7 show that model achieves better performance when fine-tuned on our dataset with corrected metadata.

## N Details of the Models

### N.1 Architecture of the Hierarchical Model

For the *hierarchical* variant of pre-trained BERT models, we use a greedy input packing strategy where we merge multiple paragraphs into one packet until it reaches the maximum of 512 tokens. We independently encode each packet of the input text using the pretrained model and obtain representations for each packet. Then we apply a non-pretrained transformer encoder to make the packet representations context-aware. Fig 7 illustrates the detailed architecture of the hierarchical model.

### N.2 Implementation Details

We use *BERT* "bert-base-uncased" (Kenton and Toutanova, 2019), and *LegalBERT* "nlpueb/legalbert-base-uncased" (Chalkidis et al., 2020) from the Transformers Hub (Wolf et al., 2020) as our

backbone models.

**Hyperparameter & Overfitting Measures:** For the hierarchical models, we employ a maximum sentence length of 128 and document length (number of sentences) of 80. The dropout rate in all layers is 0.1. We follow the hyperparameters from Chalkidis et al. 2022a with a batch size of 8 and learning\_rate of 3e-3. We train models with the Adam optimizer for up to 10 epochs. We use *PyTorch* (Paszke et al., 2019) 2.0.1.

## O Soft Loss Function

The formulation of the soft loss function is represented as

$$-\sum_{i=1}^n \sum_c p_{hum}(y_i | x_i) \log p_{\theta}(y_i = c | x_i),$$
 where we compute  $p_{hum}$  by applying a standard normalization function to the judges’ votes for each pair following Peterson et al. 2019.

## P Expected Calibration Error

Expected Calibration Error (ECE) is a measure of the difference between the predicted probabilities assigned by a model and the accuracy of those predictions. ECE is typically calculated by dividing the predicted probability space into a fixed number  $M$  of intervals (or bins) and then computing the average absolute difference between the predicted and observed probabilities within each bin  $B_m$  as follows:

$$ECE = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (1)$$

A lower ECE indicates better calibration, suggesting that the model’s predicted probabilities are more accurate reflections of the true likelihood.

## Q More Calibration Results

Tab 8 presents the calibration experiment results on both dev and test set. Soft-training improves the model’s COC performance (hm-F1) on SV cases for the development set, but lowers it for the test set. Two contributing factors account for this discrepancy: i) SV instances constitute less than 10% of the total dataset, potentially limiting the generalizability of improvements. ii) Given the temporal split of the dataset, distribution shifts in case-law may occur. For instance, legal issues addressed in SV cases could stabilize after a key legal decision. In other words, something that is legally controversial and uncertain in the training set time



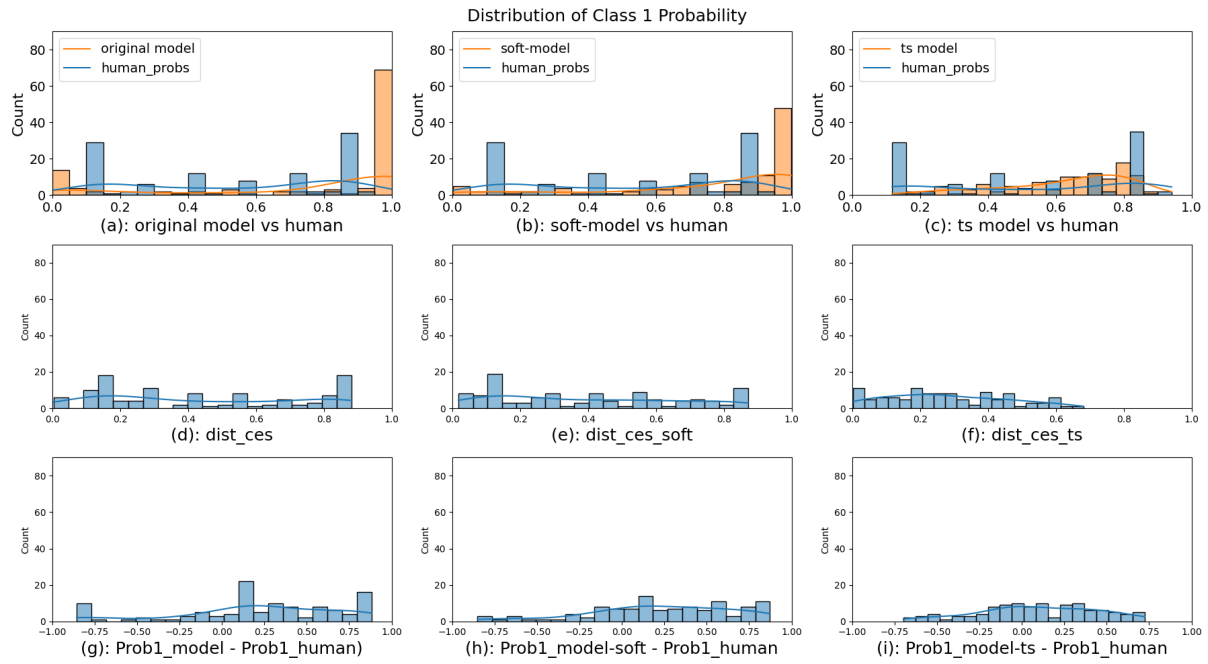


Figure 8: LegalBERT seed 1 on the SV cases in test set. Row 1: the distribution over instance-based absolute errors between probabilities for class 1 of the model vs human vote distributions. Row 2: the distribution over instance-based DistCE of models. Row 3: the distribution over errors between probabilities for class 1 of the model vs human vote distributions.

period may be settled law in the development and test dataset partitions. Our models may have overfit to the training set period. Fig 8 shows more figures comparing human uncertainty to model uncertainty.

		hm-F1			ECE			DistCE			count
		/	TS	soft	/	TS	soft	/	TS	soft	
BERT Test	u	<b>66.19 ± 3.55</b>	<b>66.19 ± 3.55</b>	65.79 ± 1.1	20.68 ± 4.08	<b>3.44 ± 1.09</b>	22.25 ± 0.36	28.35 ± 2.7	38.98 ± 0.66	<b>25.32 ± 0.42</b>	1463
	sv	<b>43.76 ± 4.23</b>	<b>43.76 ± 4.23</b>	39.73 ± 1.35	27.61 ± 5.11	<b>6.12 ± 3.8</b>	28.89 ± 0.48	39.47 ± 4.33	<b>29.73 ± 0.59</b>	42.33 ± 1.81	112
	all	<b>64.22 ± 3.46</b>	<b>64.22 ± 3.46</b>	63.7 ± 1.44	21.1 ± 4.14	<b>3.21 ± 1.4</b>	22.62 ± 0.39	29.15 ± 2.2	38.32 ± 0.58	<b>26.53 ± 0.32</b>	1575
Dev	u	68.86 ± 1.52	68.86 ± 1.52	<b>69.57 ± 0.77</b>	21.07 ± 4.72	<b>4.39 ± 1.28</b>	23.39 ± 0.62	26.85 ± 2.93	38.11 ± 0.75	<b>23.31 ± 0.72</b>	1381
	sv	41.33 ± 5.47	41.33 ± 5.47	<b>47.39 ± 4.39</b>	22.03 ± 4.34	<b>6.45 ± 1.95</b>	20.76 ± 1.49	37.44 ± 2.59	<b>26.74 ± 1.8</b>	34.53 ± 0.27	135
	all	66.49 ± 2.19	66.49 ± 2.19	<b>67.58 ± 0.53</b>	20.99 ± 4.88	<b>4.2 ± 1.47</b>	23.04 ± 0.47	27.79 ± 2.52	37.09 ± 0.84	<b>24.31 ± 0.64</b>	1516
LegalBERT Test	u	<b>69.3 ± 1.88</b>	<b>69.3 ± 1.88</b>	67.31 ± 0.5	23.32 ± 1.01	<b>2.95 ± 0.67</b>	22.04 ± 0.63	25.1 ± 1.17	37.23 ± 1.71	<b>24.7 ± 0.17</b>	1463
	sv	<b>53.67 ± 4.59</b>	<b>53.67 ± 4.59</b>	46.21 ± 4.41	29.92 ± 2.81	<b>8.02 ± 0.58</b>	28.49 ± 0.82	41.03 ± 2.75	<b>28.28 ± 0.99</b>	40.72 ± 1.0	112
	all	<b>68.03 ± 1.29</b>	<b>68.03 ± 1.29</b>	65.16 ± 0.82	23.75 ± 1.08	<b>2.99 ± 0.83</b>	22.32 ± 0.64	26.23 ± 1.0	36.6 ± 1.52	<b>25.84 ± 0.21</b>	1575
Dev	u	<b>71.12 ± 0.67</b>	<b>71.12 ± 0.67</b>	70.38 ± 0.67	24.69 ± 0.84	<b>2.33 ± 1.65</b>	23.4 ± 0.75	23.83 ± 0.29	36.21 ± 1.57	<b>22.9 ± 0.96</b>	1381
	sv	42.17 ± 2.92	42.17 ± 2.92	<b>44.51 ± 3.23</b>	24.08 ± 0.96	<b>8.48 ± 2.47</b>	24.87 ± 1.92	40.91 ± 1.2	<b>27.44 ± 0.81</b>	35.4 ± 1.63	135
	all	<b>68.29 ± 0.98</b>	<b>68.29 ± 0.98</b>	68.28 ± 0.66	24.36 ± 0.97	<b>2.35 ± 1.02</b>	23.26 ± 1.0	25.35 ± 0.3	35.43 ± 1.44	<b>24.02 ± 0.75</b>	1516

Table 8: COC performance (hm-F1), confidence-calibration (ECE), and human-calibration (DistCE) performance with std( $\pm$ ) in dev and test set. "/": COC finetuned models on § 6.1; "soft": models fine-tuned with soft-loss; "TS": model after Temperature Scaling. We choose the temperature  $t$  by searching a range of possible values for  $t$  on the dev set. We noted that the chosen  $t$  across three random seeds were consistently overly aggressive, with values of 5.5, 5.8, and 5.5. Calibration results with standard deviation; Results shown over 3 random seeds.