Fairness Beyond Performance: Revealing Reliability Disparities Across Groups in Legal NLP

Santosh T.Y.S.S, Irtiza Chowdhury
School of Computation, Information, and Technology;
Technical University of Munich, Germany
{santosh.tokala, irtiza.chowdhury}@tum.de

Abstract

Fairness in NLP must extend beyond performance parity to encompass equitable reliability across groups. This study exposes a critical blind spot: models often make less reliable or overconfident predictions for marginalized groups, even when overall performance appears fair. Using the FairLex benchmark as a case study in legal NLP, we systematically evaluate both performance and reliability disparities across demographic, regional, and legal attributes spanning four jurisdictions. We show that domain-specific pre-training consistently improves both performance and reliability, especially for underrepresented groups. However, common bias mitigation methods frequently worsen reliability disparities, revealing a trade-off not captured by performance metrics alone. Our results call for a rethinking of fairness in high-stakes NLP: To ensure equitable treatment, models must not only be accurate, but also reliably self-aware across all groups.

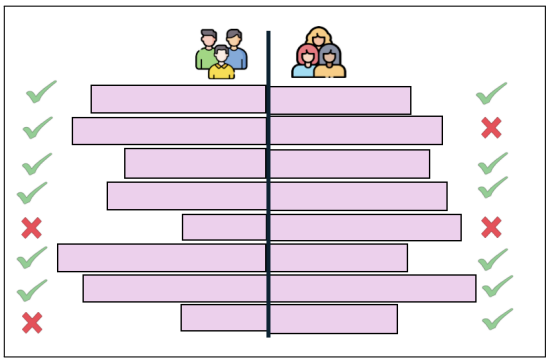

Figure 1: Toy example illustrating performance parity but reliability disparity across groups. The two groups, male and female, each contain 8 examples with 6 out of 8 correct predictions, indicating equal accuracy. Each prediction is represented by a horizontal bar, where the bar length encodes the model's confidence. Correct and incorrect predictions are marked using ✓ and ×, respectively. While accuracy is balanced across groups, the female group's incorrect predictions occur at substantially higher confidence levels, revealing overconfident mispredictions. This highlights how models can appear fair under performance metrics while masking disparities in reliability across groups.

1 Introduction

As NLP systems are increasingly deployed in high-stakes domains—such as healthcare, education, finance, and law—the need for fairness, transparency, and trust has become paramount (Chen et al., 2024; Burkart and Huber, 2021; Pressman et al., 2024). Fairness in NLP has traditionally been framed through performance parity, where models are expected to achieve comparable predictive accuracy across demographic or social groups (Mehrabi et al., 2021; Gallegos et al., 2024; Li et al., 2023). However, this view overlooks a critical dimension of responsible model behavior: reliability—a model's ability to express calibrated confidence and to abstain when uncertain (Geifman and El-Yaniv, 2017; Guo et al., 2017). In safety-critical settings, a model's confidence can be as consequential as its accuracy: overconfident or erratic predictions, especially for marginalized groups, can amplify harm and erode trust (Tran et al., 2022; Ulmer, 2024).

We argue that fairness in NLP must extend beyond performance parity to encompass equitable reliability across groups. This means not only delivering equitable outcomes, but also ensuring that models are consistently calibrated, cautious, and self-aware across different subpopulations. A system that performs well on average across groups may still exhibit overconfidence or erratic uncertainty for certain groups—leading users to overtrust unreliable outputs and creating hidden disparities in how risk is experienced and mitigated.

To investigate this broader fairness paradigm, we examine the legal domain—a representative high-stakes setting, where fairness and reliability are not just technical ideals but also legal obli-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 24376–24390

gations. Legal NLP has gained attention for its potential to transform the legal field by enabling efficient, scalable analysis of complex texts, supporting tasks such as legal research, contract review, and case summarization, and reducing traditional workflow costs (Zhong et al., 2020; Ashley, 2018; Katz et al., 2023; Ariai and Demartini, 2024; Kalamkar et al., 2021). By providing data-driven tools, these systems aim to democratize access to legal insights and enhance decision-making for judges, lawyers, and policymakers (Santosh et al., 2024a,c; Chalkidis et al., 2021b; Niklaus et al., 2023).

However, as such models are integrated into sensitive legal workflows, concerns over equity intensify. In law, fairness is rooted in principles of equality and non-discrimination[1], demanding that automated systems perform consistently across different demographic groups, jurisdictions, and case types (Chalkidis et al., 2022). Models trained on historical legal data risk replicating or even amplifying entrenched biases, undermining the legitimacy of legal systems and public trust, contravening the core legal commitment to impartiality and equal treatment under the law. To address this, Chalkidis et al. (2022) introduced FairLex, a benchmark that evaluates fairness across four legal jurisdictions by measuring performance disparities across legal and demographic attributes. Separately, Santosh et al. (2024b) proposed a selective prediction framework for legal NLP, allowing models to abstain in uncertain scenarios and highlighting the importance of calibrated decision-making. Despite these advances, fairness and reliability have largely been treated as separate concerns.

In this work, we unify these perspectives by advocating for fairness through reliability. We propose that models should not only perform equitably across groups, but also exhibit uniform reliability—providing calibrated confidence and deferral behavior consistently across all subpopulations. Disparities in reliability can expose certain groups to greater risk, even when accuracy appears balanced. For example, if a model is more overconfident for specific demographic groups, its errors may be harder to detect, leading to unequal downstream consequences. To investigate this broader notion of fairness, we conduct a large-scale empirical study using the FairLex benchmark. Our goals are twofold: (i) to examine how standard fine-tuning and domain-specific pretraining affect performance and reliability disparities across groups and (ii) to evaluate the extent to which popular bias mitigation techniques improve overall performance, reliability and promote equity across groups, not only in outcomes, but also in reliability.

Our findings reveal that: (a) Domain-specific pretraining improves both performance and reliability, particularly benefiting worse-off groups and reducing disparity gaps.(b) Representational imbalance and distributional misalignment partially explain performance disparities but fail to account for reliability gaps, highlighting the need to investigate their distinct causes. (c) High performance does not imply better reliability or calibration; in fact, higher-performing groups often exhibit overconfidence on incorrect predictions, undermining trust. (d) Bias mitigation methods, while occasionally improving performance parity, frequently exacerbate reliability disparities, revealing trade-offs that accuracy-based fairness evaluations fail to capture.(e) No single method consistently improves all fairness dimensions across settings, underscoring the need for context-aware approaches that jointly address both disparities.

Our results suggest that reliable model behavior, particularly in how uncertainty is handled across groups, must be a core component of fairness evaluation in NLP, especially in high-stakes applications. While our study focuses on legal NLP as a case study, the insights are broadly applicable and call for a shift in how fairness is conceptualized and operationalized across the NLP landscape.

2 Related Work

Uncertainty Quantification & Selective Prediction Uncertainty is a core concept in machine learning, reflecting a model's confidence in its predictions. In selective prediction, a model can abstain from predicting when uncertainty is high—allowing it to defer decisions it deems unreliable. This area has been a focus of research (Chow, 1957; Hellman, 1970; Fumera and Roli, 2002; Cortes et al., 2016; El-Yaniv et al., 2010; Geifman and El-Yaniv, 2017) and has gained attention in NLP tasks, such as question answering (Kamath et al., 2020; Garg and Moschitti, 2021), classification (Gu and Hopkins, 2023; Varshney et al., 2022b,a), knowledge probing (Yoshikawa and Okazaki, 2023), and text generation (Ren et al.,

[1]The legal notion of discrimination differs in scope and semantics from the concepts of fairness and bias in machine learning (Chalkidis et al., 2022; Gerards and Xenidis, 2021).

2022; Chen et al., 2023; Cole et al., 2023; Wen et al., 2024). A related area, though distinct is model calibration (Jiang et al., 2018; Desai and Durrett, 2020; Wang et al., 2020a; Guo et al., 2017), which aims to develop interpretable confidence measures that adjust a model's overall confidence level, unlike selective prediction which focuses on relative confidence across individual examples.

Uncertainty quantification techniques originally developed for classification and regression in deep learning (Gal et al., 2016) have been adapted for NLP tasks using encoder-only language models like BERT (Zhang et al., 2019; He et al., 2020; Shelmanov et al., 2021; Vazhentsev et al., 2022; Kotelevskii et al., 2022; Wang et al., 2022; Xin et al., 2021). They have been extended to text generation tasks, which is more challenging, as it requires estimating uncertainty for entire sequences, which can involve an exponential or infinite number of predictions (Malinin and Gales, 2020; Kuhn et al., 2023; Ren et al., 2022; Vazhentsev et al., 2023; Lin et al., 2023; Van der Poel et al., 2022; Vashurin et al., 2024; Fadeeva et al., 2023, 2024).

In legal NLP, Santosh et al. (2024b) explored factors such as pre-training corpus, model size, confidence estimator, and loss function on reliability in case outcome classification tasks using selective prediction. Our work extends this line by examining group-level disparities in reliability and how bias mitigation methods impact both performance and reliability equity. We argue that fairness evaluations should include not just equitable outcomes, but also consistent reliability across groups, ensuring that all subpopulations receive dependable predictions and are equally protected from overconfident errors.

Fairness Fairness in machine learning addresses various forms of discrimination (Makhlouf et al., 2021; Mehrabi et al., 2021), including group fairness, individual fairness, and causality-based fairness. Group fairness ensures equitable predictions across demographic subgroups, avoiding differential treatment based on attributes like race, gender, or age (Hardt et al., 2016; Zafar et al., 2017; Corbett-Davies et al., 2017). Individual fairness focuses on treating similar individuals similarly (Sharifi-Malvajerdi et al., 2019; Yurochkin et al.; Dwork et al., 2012), while causality-based fairness aims to reduce biases from confounding variables (Wu et al., 2019; Zhang and Bareinboim, 2018). To address fairness and bias issues, several bias mitigation methods are proposed which can be categorized into pre-processing (modifying data before training) (Kamiran and Calders, 2012; Calmon et al., 2017), in-processing (adjusting the model during training) (Kamishima et al., 2012; Beutel et al., 2017; Zhang et al., 2018; Mehrotra and Vishnoi, 2022; Madras et al., 2018), and post-processing (altering outputs after training) (Hardt et al., 2016; Jiang et al., 2020).

In the legal domain, fairness research is emerging due to the societal consequences of biased decisions. Angwin et al. (2016) identified racial bias in the COMPAS parole risk assessment tool and Wang et al. (2021) found gender disparities in legal judgment prediction. Chalkidis et al. (2022) developed the FairLex benchmark to assess bias mitigation algorithms across different groups. While most work focuses on group fairness, recent work explores counterfactual fairness in lower court decisions (Santosh et al., 2024a).

In our study, we adopt the group fairness criterion and examine in-processing bias mitigation techniques, as benchmarked in FairLex. Unlike prior work, we analyze not just whether performance disparities are reduced, but also whether reliability is equitably maintained across all groups. By unifying uncertainty estimation and fairness, we aim to expand the fairness agenda to account for how models express and manage uncertainty—through calibrated confidence and appropriate abstention across groups—particularly in high-stakes settings.

3 Tasks & Datasets

We describe the four datasets used in Fairlex (Chalkidis et al., 2022), along with their tasks, associated groups under each attribute.

ECHR (Chalkidis et al., 2021a) contains 11,000 cases from the European Court of Human Rights, which adjudicates complaints against states for alleged human rights violations. Given case facts, the task is to predict the set of articles deemed violated by the court, making this a multi-label classification across 10 articles. The data is chronologically split into 9,000 cases for training (2001–16), 1,000 for development (2016–17), and 1,000 for testing (2017–19). Attributes include (a) defendant states (grouped into Central-Eastern or other European states), (b) applicant's age group ($\leq$ 35, 35–64, or $\geq$ 64), and (c) applicant's gender (male or female).

SCOTUS consists of 9,262 cases from US Supreme Court opinions, which generally address

only the most complex or controversial cases unresolved by lower courts. The task is to predict the issue area based on the court opinion, a single-label multi-class classification task with 14 labels. The cases are split as follows: 7,400 for training (1946–1982), 914 for development (1982–1991), and 931 for testing (1991–2016). Attributes include (a) the type of respondent (categorized as person, public entity, organization, or facility) and (b) decision direction (liberal or conservative).

FSCS (Niklaus et al., 2021) includes 85,000 decisions from the Federal Supreme Court of Switzerland, which addresses complex cases unresolved by lower courts and are available in one of three languages (German, French or Italian). Given the facts of case, the task is to predict whether the plaintiff's request is approved or dismissed, a binary classification. Cases are chronologically split into train (59.7k, 2000–2014), development (8.2k, 2015–2016), and test (17.4k, 2017–2020) sets. Attributes include (a) Decision language (German, French, or Italian), (b) legal area (public law, penal law, social law, civil law, or insurance law), and (c) originating canton region (R. Lémanique, Zürich E., Mittelland, E. Switzerland, N.W. Switzerland, C. Switzerland, Ticino or Federation).

CAIL (Xiao et al., 2018; Wang et al., 2021) contains over 1 million cases from the Supreme People's Court of China, originating from high courts on matters of national importance. The dataset includes labels for criminal code article prediction, charge type, imprisonment term, and monetary penalty. Following Chalkidis et al. (2022), the imprisonment term prediction is reframed as a crime severity classification task—a single-label, multi-class task in Chinese based on case facts. This task classifies severity into 6 clusters (0, $\leq$ 12, $\leq$ 36, $\leq$ 60, $\leq$ 120, >120 months). Cases are chronologically split into 80,000 for training (2013–2017), 12,000 for development (2017–2018), and 12,000 for testing (2018). Attributes include (a) applicant's gender (male or female) and (b) the provincial-level administrative region (Beijing, Liaoning, Hunan, Guangdong, Sichuan, Guangxi or Zhejiang).

We report performance on each dataset using macro-F1 scores to account for label imbalances.

4 Background & Methods

We introduce a selective prediction framework, along with metrics for evaluating reliability and bias mitigation algorithms, we apply in our study.

4.1 Selective Prediction

A standard classifier learns a function $f : X \to Y$ that maps input X to labels Y. We extend this with a selection function $g : X \to \{0, 1\}$, forming a selective classifier $h = (f, g)$; $h \to Y \cup \{\perp\}$, where $\perp$ represents abstention. For input x, the selective classifier returns $f(x)$ when the selection function signals prediction, and abstains otherwise.

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

A common approach for g is to rely on a confidence estimator $\tilde{g} : X \to \mathbb{R}$ that assigns confidence values to instances along with a threshold $\gamma \in \mathbb{R}$. Ideally, $\tilde{g}(x)$ should be high when $f(x)$ is correct and low when incorrect, to obtain a reliable classifier with better self-awareness. We use the most widely used Softmax Response (SR) (Hendrycks and Gimpel, 2016) as confidence estimator, based on the maximum probability assigned to one of the labels:

$$g(x) = \mathbb{1}[\tilde{g}(x) > \gamma] \qquad \tilde{g}(x) = \max_{y \in Y} p(y)$$

Metrics for Selective Prediction Coverage (C) is the proportion of instances the model chooses to predict and risk (R) is the error on those subset of predictions. For selective classifier $h = (f, g)$ on dataset D with inputs x_i and true labels y_i:

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

$$R(h) = \frac{\frac{1}{|D|} \sum_{(x_i, y_i) \in D} l(f(x_i), y_i) \cdot g(x_i)}{C(h)}$$

where loss function l measures the error between prediction $f(x_i)$ and ground truth y_i. A reliable model should achieve high coverage at low risk, meaning accurate predictions for many instances and abstentions on incorrect ones. Lowering γ increases coverage but may also raise risk, leading to a risk-coverage trade-off. Thus, we plot coverage against risk to create a Risk-Coverage curve (El-Yaniv et al., 2010) and calculate the Area under Risk-Coverage Curve (AuRCC), where a lower AuRCC indicates a better reliable classifier.

Xin et al. (2021) propose Reversed Pair Proportion (RPP), a normalized form of the Kendall-Tau distance, to assess deviation from ideal case where $\tilde{g}$

should rank all correct predictions above all incorrect ones. RPP measures the proportion of instance pairs with a reversed confidence-error relationship. Gu and Hopkins (2023) propose a refinement metric that normalizes by the worst-case Kendall-Tau distance, yielding a calibrated, interpretable score where values of 0, 0.5, and 1 indicate optimal, random, and worst cases, respectively:

$$Rf = \frac{\sum_{1 \leq i,j \leq |D|} \mathbb{1}[\tilde{g}(x_i) < \tilde{g}(x_j), l_i < l_j]}{c(|D| - c)}$$

where c denote number of correct predictions made by the prediction function. While, AuRCC measures threshold-sensitive trade-off between risk and coverage, Rf covers confidence ranking across all instances. Thus, AuRCC depends on the separability of confidence scores while Rf depends on the monotonicity of the confidence-error relationship.

4.2 Fairness

Consider a labeled dataset with each instance consisting of an input x_i, a target label y_i and an associated group attribute $g_i \in A$ (e.g., Under attribute gender A, group g_i could be male or female). To evaluate fairness, in line with prior work (Chalkidis et al., 2022), fairness is defined in terms of performance parity (Hashimoto et al., 2018), where a fair classifier demonstrates similar performance across all groups in A. To quantify performance disparities, we use the Group Disparity, which captures the variance in macro-F1 scores across groups:

$$\text{GD} = \sqrt{\frac{1}{|G|} \sum_{g \in G} (\text{mF1}_g - \overline{\text{mF1}})^2}$$

where mF1_g represents macro-F1 score for group g, and $\overline{\text{mF1}}$ is average macro-F1 score across all groups. A higher GD value indicates that certain groups exhibit disproportionately lower model performance compared to others. In addition to assessing performance disparities, we argue it is crucial to evaluate how the reliability of predictions varies across groups. Thus we can extend the Group Disparity metric to the Reliability measures, such as AuRCC or Refinement in place of macro-F1, emphasizing the need for both equal performance and consistent reliability across all groups.

Bias Mitigation Methods Traditional fine-tuning methods optimize the loss equally over all instances in a given batch, referred to as Empirical Risk Minimization (ERM) (Vapnik, 1991). We also utilize a variant of ERM with a balanced group sampler (GS), where an equal number of instances from each group are included in each batch. We explore several bias mitigation methods derived from group-robust or domain-invariant representation learning, which enhance transferability across groups by eliminating group-specific information and thus prevent overfitting to specific groups and improve generalization across groups.

GDRO (Sagawa et al., 2019) Group Distributionally Robust Optimization aims to optimize the worst-group loss wherein the each group-wise aggregated losses are weighted inversely proportional to the performance of instances in that group.

V-REx (Krueger et al., 2021) hypothesize that variation of training loss across groups is representative of the variation later encountered at test time, so they also consider the variance across the group-wise losses as an extra loss term.

IRM (Arjovsky et al., 2019) seeks to estimate non-linear, invariant, causal predictors from multiple training environments, to ensure better generalization. To learn invariances across environments, they fit an optimal classifier on top of that representation matches for all environments.

DeepCORAL (Sun and Saenko, 2016) aligns the second-order statistics of the distributions across groups by penalizing the differences in covariance between the feature distributions of each pair of groups, to obtain group invariant representations.

Adversarial Removal (Elazar and Goldberg, 2018) uses an additional adversarial classifier to remove the group-specific attribute information encoded in the feature representations. This classifier is applied on top of the feature extractor to predict the group for each instance. Through adversarial training, we maximize the feature extractor's ability to capture relevant information for the main classification task while minimizing its ability to predict the group (Santosh et al., 2022).

5 Experiments

Models: Following earlier works of Chalkidis et al. (2021b, 2022), we use a hierarchical BERT-based model to account for the longer input lengths. We utilize a pre-trained BERT model to encode each paragraph in the input independently, obtaining the [CLS] token representation for each paragraph, which is then passed to a two-layer transformer (similar to the configuration of the first layer) to obtain context-enriched representations. These are

			ERM-DPT			ERM-Gen		
	% t	KLd	mF1 ↑	AuR ↓	Rf ↓	mF1 ↑	AuR ↓	Rf ↓
ECHR								
Applicant Age								
<= 35	14	0.19	42.90	1.67	25.94	40.27	1.46	23.06
<= 65	**68**	**0.18**	**51.00**	6.99	**9.03**	**51.13**	7.19	**10.47**
>65	18	0.32	49.07	**1.33**	20.65	44.03	**1.45**	17.91
Applicant Gender								
Male	**77**	**0.17**	**48.20**	7.53	**10.39**	**48.44**	7.70	**12.54**
Female	23	0.26	45.10	**1.41**	12.51	43.87	**1.55**	18.23
Defendant State								
E.C	**80**	**0.17**	**50.27**	13.44	**10.33**	**49.87**	14.47	**11.90**
West	20	0.28	43.77	**1.88**	27.58	41.37	**1.92**	27.63
SCOTUS								
Decision Direction								
Liberal	**52**	**0.04**	**84.27**	24.69	19.96	**81.23**	29.73	20.09
Conservative	48	0.05	81.77	**23.98**	**15.62**	76.43	**28.22**	**17.55**
Respondent Type								
Person	34	**0.05**	80.10	19.17	20.01	75.83	24.00	20.91
Organization	13	0.11	88.50	5.93	21.86	86.23	7.45	23.93
Public Entity	**51**	0.07	81.50	18.38	**16.26**	74.13	22.83	**18.24**
Facility	3	0.26	**89.07**	**1.40**	26.03	**89.03**	**1.54**	24.37
FSCS								
Court Region								
R. Léman.	**27**	0.04	71.30	382.72	25.33	69.75	601.29	32.51
Zürich	18	0.04	68.17	168.08	20.19	68.21	222.36	24.54
E. Mittel.	17	0.08	70.27	130.72	19.57	67.16	215.71	24.32
N.W. Swiss	11	0.02	**71.77**	94.45	**17.19**	**70.43**	134.96	21.50
E. Swiss	12	0.03	71.20	105.43	22.40	69.61	135.05	25.77
C. Swiss	10	0.03	67.93	80.24	21.39	66.34	104.17	23.49
Ticino	6	**0.00**	62.70	**33.69**	18.15	64.42	**50.89**	**20.45**
Federation	3	**0.00**	62.10	114.94	33.28	64.14	151.15	38.57
Decision Language								
German	**60**	**0.03**	68.80	646.57	20.91	67.92	877.39	24.53
French	35	**0.03**	**71.40**	485.48	24.95	**69.34**	783.50	32.13
Italian	5	0.04	64.10	**40.81**	**18.50**	63.71	**70.14**	**22.93**
Legal Area								
Public Law	**31**	**0.00**	56.53	741.66	33.27	55.93	1,200.32	42.32
Penal Law	25	**0.00**	**82.13**	212.57	20.18	**81.64**	**243.49**	**21.63**
Social Law	24	0.02	64.93	**197.37**	21.73	65.91	309.05	28.26
Civil Law	20	0.06	70.10	212.61	**18.45**	68.63	360.10	24.86
CAIL								
Court Region								
Beijing	**21**	**0.05**	**65.63**	**140.39**	**22.77**	**60.15**	**204.00**	**25.62**
Liaoning	17	**0.05**	56.70	408.93	33.81	46.75	517.66	36.98
Hunan	16	**0.05**	58.97	404.85	33.17	54.45	517.18	35.39
Guangdong	15	**0.05**	57.60	262.24	30.14	49.85	353.61	32.51
Sichuan	14	0.06	56.03	338.40	33.76	51.20	424.67	36.18
Guangxi	11	0.07	60.47	306.71	34.42	50.75	376.27	35.99
Zhejiang	5	0.07	56.27	263.96	32.60	54.95	351.77	36.55
Defendant Gender								
Male	**92**	**0.03**	59.13	1886.90	31.36	**53.60**	2459.77	33.99
Female	8	0.08	**59.60**	**151.24**	**30.01**	51.00	**200.25**	**33.16**

Table 1: Comparison of model performance (mF1) and reliability measures (AuRCC, Rf) for fine-tuned models using ERM, initialized with either domain-specific pretrained (DPT) or general-purpose (Gen) models, across groups defined by each attribute on four datasets. Representation inequality and temporal concept drift are quantified using %t (percentage of the training split) and KLd (KL-Divergence), respectively. AuR represents AuRCC.

max-pooled to obtain the final representation of the input, which is sent to the classification layer.

We use the four mini-sized BERT models from (Chalkidis et al., 2022), which are continually pre-trained on the corpora of these four datasets, as our initialization in the first-level hierarchical model. We also experiment with their base models, MiniLMv2 model checkpoints (Wang et al., 2020b), which is the distilled version of RoBERTa (Liu, 2019) for the English datasets (such as ECHR and SCOTUS), and the one distilled from XLM-R (Conneau, 2019) for the rest (trilingual FSCS, and Chinese CAIL). We run each experimental variant with five random seeds to report mean results. Implementation details are provided in App. A.

5.1 Disparities across Groups with ERM

We present the performance (mF1) and reliability measures (AuRCC, Rf) for each group under each attribute across four datasets, using ERM fine-tuned with domain-specific pre-trained and general models as initializations, in Table 1. Additionally, we analyze factors related to data distribution for each group to study their relationships: (a) Representation Inequality: The extent to which each group is represented in the training data, measured by the percentage of training instances belonging to the group. (b) Temporal Concept Drift: Changes in label distributions over time for each group, due to the chronological nature of dataset splits, quantified using the KL divergence between training and test label distributions.

ECHR Groups with higher representation in the training split and lower KL divergence (e.g., <=65 under the attribute applicant age) achieve higher macro-F1 scores, likely due to better alignment with the training distribution, which aids generalization. However, these groups exhibit higher AuRCC values, indicating overconfident predictions, as the model struggles to abstain from making incorrect predictions. This highlights that high performance does not necessarily equate to well-calibrated or reliable decision-making. Notably, lower Rf values in high-performing groups suggest effective ranking of predictions (correct versus incorrect), but this does not align with threshold-based AuRCC, emphasizing a disconnect between these metrics. Across all attributes (age, gender, state), we observe stark disparities in macro-F1, AuRCC, and Rf values among different groups, with pronounced differences in reliability metrics. Domain-specific pretraining improves both reliabil-

ity and performance across groups while narrowing disparities, especially benefiting underrepresented groups, compared to general models.

SCOTUS For decision direction, better-represented groups exhibit higher performance but slightly worse reliability, with disparities being less pronounced due to more balanced data distributions. Interestingly, the least-represented group with the highest KL divergence under respondent type (i.e., Facility) achieves better performance and AuRCC, suggesting that representation is not always the sole determinant of reliability or performance. A positive correlation between AuRCC and Rf is observed for decision direction, contrasting with the inverse trend in respondent type, indicating that group-specific factors influence reliability differently. Consistently, domain-specific pretraining improves both performance and reliability across groups, further reducing disparities.

FSCS Representation inequality partially explains performance disparities in decision language and court region but not in legal area. Interestingly, lower-performing groups often exhibit better reliability in decision language and court region. For instance, Public Law, despite high representation and no KL divergence, shows the lowest performance and the worst reliability (AuRCC, Rf). One possible explanation is the more consistent judicial outcomes in Penal Law, which align better with model predictions, as noted in prior studies (Niklaus et al., 2021; Chalkidis et al., 2022). Domain-specific pretraining consistently enhances performance and reliability across all groups.

CAIL In the gender attribute, males exhibit higher AuRCC than females, despite similar performance and Rf values, indicating poorly calibrated confidence scores for male instances. For region, lower-performing groups tend to have higher Rf and AuRCC values, reflecting reduced reliability. Domain-specific pretraining improves both performance and reliability, while reducing disparities across groups.

Key Takeaways (i) Highly representative groups closer to the training distribution typically achieve better performance but often exhibit overconfident predictions, evidenced by higher AuRCC values. (ii) High ranking quality (Rf) does not always align with reliable abstention (AuRCC), particularly for high-performing groups. (iii) While low-performing groups often exhibit better reliability (lower AuRCC, Rf), exceptions underscore calibration challenges, emphasizing the need for further investigation into underlying causative factors. (iv) Domain-specific pretraining consistently enhances both performance and reliability, reducing disparities across groups compared to general models. (v) Certain group attributes (e.g., respondent type in SCOTUS or public law in FSCS) deviate from typical trends, highlighting the need for deeper analysis of data characteristics and task complexity.

5.2 Performance of Bias Mitigation methods

Since these methods are applied to groups under each specific attribute, we evaluate each of the bias mitigation method across that specific attribute. Fig. 2 presents the average macro-F1 (mF1) across groups, reliability measures (AURCC, RF), and group disparity metrics (calculated for mF1, AURCC, and RF) across four datasets, using domain-specific pre-trained model initializations. Detailed results are provided in App. B.

ECHR *Applicant age:* Balanced Group Sampler (ERM+GS) not only improves performance but also enhances reliability metrics while significantly reducing disparities in both performance and reliability (AuRCC, Rf). DeepCoral also mitigates performance disparity, though at the expense of overall performance. But it increases AuRCC and Rf, as well as their respective disparities, highlighting the complex relationship between performance and reliability. GDRO and V-REx improve Rf and reduce Rf disparity but do not neither improve AuRCC or performance, nor their disparities. *Applicant Gender:* ERM+GS continues to improve performance and reliability, while also reducing AuRCC and performance disparities. However, it struggles to improve Rf. AdvRem reduces Rf disparity but worsens overall Rf, while GDRO improves AuRCC and reduces its disparity but underperforms in terms of Rf. *Defendant State:* ERM+GS delivers improvements in both performance and reliability, while also reducing disparities. Other algorithms show modest performance gains but fail to reduce disparities and experience declines in both the reliability metrics. GDRO and V-REx reduce Rf disparity but at the expense of overall Rf.

SCOTUS *Decision Direction:* IRM stands out by enhancing performance and reducing both performance and Rf disparities, though it sacrifices overall Rf. GDRO narrows disparities in AuRCC and Rf, but with a slight decline in AuRCC. Other methods, including ERM+GS, fail to reduce performance disparity even with reduced overall performance. AdvRem, GDRO, and DeepCoral improve

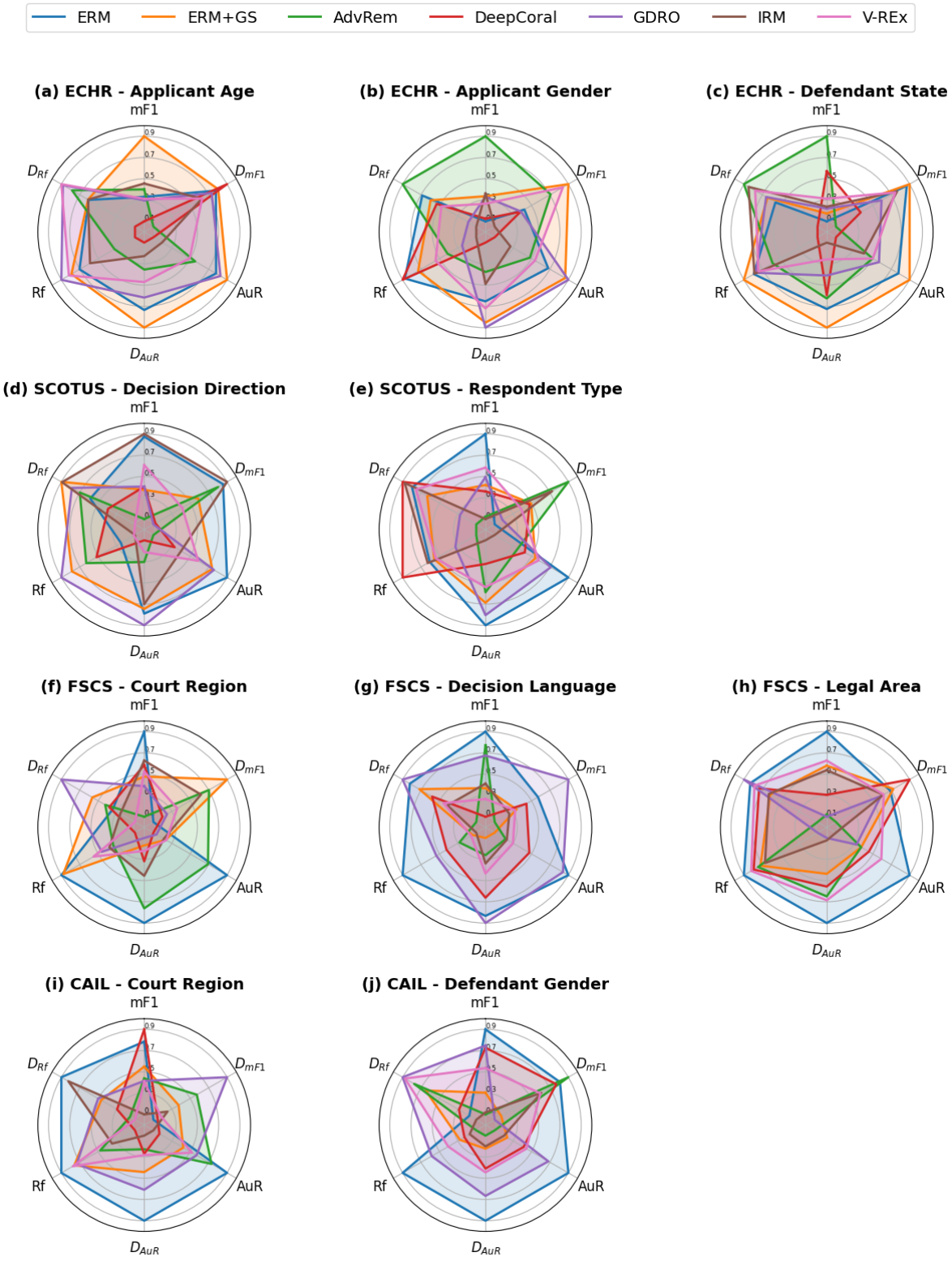

Figure 2: Results of bias mitigation algorithms applied to domain-specific pre-trained models for each attribute in the ECHR, SCOTUS, FSCS and CAIL datasets. We report the average macro-F1 score across groups (mF1), reliability metrics (AuRCC (AuR), Rf), and group disparities in performance (D_{mF1}), AuRCC (D_{AuR}), and refinement (D_{Rf}). To ensure consistent interpretation, metrics where lower is originally better (AuR, Rf, and disparity metrics) have been inverted so that higher values indicate better performance. The macro-F1 score (mF1) is reported as is since higher is naturally better. Therefore, for all metrics shown, higher scores correspond to better outcomes.

overall Rf while decreasing Rf disparity. *Respondent Type:* All methods reduce performance disparity, but at the cost of overall performance. AdvRem achieves the best improvement in performance disparity but with the greatest drop in overall performance. None of the methods improve AuRCC or its disparity. Only DeepCoral, IRM improve overall Rf while reducing Rf disparity.

FSCS *Court Region:* All methods narrow performance disparity, with ERM+GS performing the best, albeit with a slight loss in overall performance. None improve overall AuRCC or its disparity. While overall Rf remains unimproved, ERM+GS, GDRO, AdvRem, and DeepCoral reduce Rf disparity. *Decision Language:* GDRO achieves slight overall performance improvement while narrowing disparities in performance, AuRCC, and Rf. Other methods do not perform as well. *Legal Area:* ERM+GS and DeepCoral reduce performance disparity but at the expense of overall performance. None of the methods improve AuRCC, Rf, or their disparities.

CAIL *Court Region:* All methods reduce performance disparity but exhibit overall performance losses (except DeepCoral). None improve AuRCC, Rf, or their disparities. *Gender:* Only AdvRem narrows performance disparity, but this comes at the cost of overall performance. Most methods, except IRM, narrow Rf disparity but at the expense of overall Rf. GDRO strikes a better balance, achieving a trade-off between overall reliability and disparity reduction for both AuRCC and Rf.

Key Takeaways: While most bias mitigation methods show partial success in reducing performance disparities, they often fail to maintain well-calibrated predictions, leading to widening disparities in reliability metrics. This highlights a critical gap: achieving equity in performance does not guarantee equitable reliability across groups. Therefore, future research must focus on developing methods that simultaneously address both performance and reliability disparities. No single method consistently improves all key metrics—performance, AuRCC, Rf, and their disparities—across all datasets. This suggests that bias mitigation approaches need to be tailored to the unique characteristics of each dataset and fairness objective. Among the evaluated methods, Balanced Group Sampling (ERM+GS) was particularly effective in reducing disparities in both performance and reliability metrics, especially in the ECHR dataset. While GDRO showed promise by improving overall AURCC and narrowing AuRCC disparity across groups, IRM demonstrated potential in improving Rf and reducing its disparity.

6 Conclusion

This work examined fairness in legal NLP through the lens of reliability, challenging the assumption that equitable performance alone ensures equitable treatment. Using the FairLex benchmark, we showed that models can exhibit significant disparities in confidence calibration and abstention behavior across groups—even when performance appears balanced. Such reliability gaps pose subtle but serious risks in high-stakes legal settings, where overconfident or erratic predictions can undermine trust, compound harm, or evade accountability. Our findings demonstrate that domain-specific pretraining improves both performance and reliability, particularly for worse-off groups. In contrast, popular bias mitigation methods, though sometimes effective at reducing performance gaps, frequently worsen reliability disparities—highlighting the limitations of current fairness interventions that focus solely on outcome parity. These results suggest the need to reframe fairness as a multi-dimensional objective, where reliability becomes a central concern alongside accuracy. In legal NLP, this means building models that are not just accurate, but also transparently uncertain, well-calibrated, and cautious in ambiguous or sensitive contexts. Such properties are essential not only for technical robustness but also for enabling meaningful human oversight, contestability, and trust in sociolegal applications. While our study centers on legal NLP, the insights extend to other high-stakes domains—such as healthcare, education, and finance—where decisions informed by NLP models carry real-world consequences. Future work should investigate how fairness-aware objectives can jointly optimize performance and reliability; how evaluation metrics can better capture disparities in model uncertainty; and how deployment practices (e.g., abstention, escalation to human review) can be designed to serve all groups equitably.

This work invites the community to reflect on a deeper question: Can a model truly be fair if it is confident for some and uncertain for others? As NLP systems increasingly inform decisions with real-world consequences, fairness must not only concern what a model predicts, but also how, when, and for whom it chooses to speak or remain silent.

Limitations

While the FairLex benchmark spans multiple jurisdictions, languages, and attributes, it may not fully represent the diversity of real-world legal systems. Certain legal contexts or demographic groups might be underrepresented, potentially limiting the generalizability of our findings. Our analysis focuses on reliability using selective prediction but does not exhaustively explore other aspects of reliability, such as robustness to adversarial examples or interpretability, which are also crucial in high-stakes legal contexts. Our study evaluates disparities across predefined single attributes such as demographics, jurisdictions and legal domains. However, these attributes may not capture the full complexity of intersectional identities that might affect both the performance and the reliability.

Ethics Statement

In conducting this research, we used publicly available datasets from the FairLex benchmark, curated by Chalkidis et al. (2022), with attention to ethical considerations. These datasets were designed to assess fairness across multiple jurisdictions, languages, and demographic attributes, providing a broad foundation for evaluating performance and reliability disparities. However, it is important to note that, as stated in Fairlex, some protected attributes, such as gender and age in the ECtHR dataset, were extracted automatically using regular expressions or manually clustered by the authors (e.g., defendant state in ECtHR dataset and respondent attribute in SCOTUS dataset). These simplifications, such as binarizing gender, may not be appropriate in real-world applications, where more nuanced categorizations are needed. Additionally, the datasets' ground truth (with the exception of SCOTUS) reflects the subjective interpretation of judges within specific jurisdictions (e.g., ECHR, Swiss, Chinese), and as such, the labels can be subjective, especially for non-trivial cases.

This study aims to address disparities in both performance and reliability in Legal NLP systems, with the goal of advancing fairness and trustworthiness in high-stakes legal applications. We recognize the significant societal and ethical implications of deploying such systems, especially in the legal domain, where biases and disparities could perpetuate systemic inequities, potentially impacting individuals' rights and access to justice. By systematically investigating disparities across attributes like demographic groups, jurisdictions, and legal domains, our research seeks to highlight and mitigate these risks. Moreover, we emphasize the ethical importance of addressing reliability disparities, as uneven reliability across groups can undermine trust and deepen inequities in decision-making processes. By prioritizing uniform reliability alongside performance, we advocate for the development of Legal NLP systems that treat all groups equitably, ensure consistent model behavior, and promote transparency and accountability. This commitment aligns with our broader goal of supporting ethical, equitable, and socially responsible advancements in Legal NLP.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23:77–91.

Farid Ariai and Gianluca Demartini. 2024. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *arXiv preprint arXiv:2410.21306*.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Kevin D Ashley. 2018. Automatically extracting meaning from legal texts: opportunities and challenges. *Ga. St. UL Rev.*, 35:1117.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021a. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. *arXiv preprint arXiv:2103.13084*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021b. Lexglue: A

benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Felix Schwemer, and Anders Søgaard. 2022. 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*.

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. 2024. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*.

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*.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806.

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.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

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

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. 2023. Lm-polygraph: Uncertainty estimation for language models. *arXiv preprint arXiv:2311.07383*.

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 et al. 2016. Uncertainty in deep learning.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

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.

Janneke Gerards and Raphaële Xenidis. 2021. *Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law*. European Commission.

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.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without

demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR.

Jianfeng He, Xuchao Zhang, Shuo Lei, Zhiqian Chen, Fanglan Chen, Abdulaziz Alhamadani, Bei Xiao, and ChangTien Lu. 2020. Towards more accurate uncertainty estimation in text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8362–8372.

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*.

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.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. 2020. Wasserstein fair classification. In *Uncertainty in artificial intelligence*, pages 862–872. PMLR.

Prathamesh Kalamkar, Janani Venugopalan, and Vivek Raghavan. 2021. Benchmarks for indian legal nlp: a survey. In *JSAI International symposium on artificial intelligence*, pages 33–48. Springer.

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.

Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. 2023. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.

Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, Fedor Noskov, Alexander Fishkov, Artem Shelmanov, Artem Vazhentsev, Aleksandr Petiushko, and Maxim Panov. 2022. Nonparametric uncertainty quantification for single deterministic neural network. *Advances in Neural Information Processing Systems*, 35:36308–36323.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5):102642.

Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Anay Mehrotra and Nisheeth Vishnoi. 2022. Fair ranking with noisy protected attributes. *Advances in Neural Information Processing Systems*, 35:31711–31725.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. *arXiv preprint arXiv:2110.00806*.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. *arXiv preprint arXiv:2301.13126*.

Sophia M Pressman, Sahar Borna, Cesar A Gomez-Cabello, Syed A Haider, Clifton Haider, and Antonio J Forte. 2024. Ai and ethics: a systematic review of the ethical considerations of large language model use in surgery research. In *Healthcare*, volume 12, page 825. MDPI.

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*.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

TYS Santosh, Kevin D Ashley, Katie Atkinson, and Matthias Grabmair. 2024a. Towards supporting legal argumentation with nlp: Is more data really all you need? *arXiv preprint arXiv:2406.10974*.

TYS 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. *arXiv preprint arXiv:2210.13836*.

TYSS Santosh, Irtiza Chowdhury, Shanshan Xu, and Matthias Grabmair. 2024b. The craft of selective prediction: Towards reliable case outcome classification–an empirical study on european court of human rights cases. *arXiv preprint arXiv:2409.18645*.

TYSS Santosh, Cornelius Weiss, and Matthias Grabmair. 2024c. Lexsumm and lext5: Benchmarking and modeling legal summarization tasks in english. *arXiv preprint arXiv:2410.09527*.

Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. 2019. Average individual fairness: Algorithms, generalization and experiments. *Advances in neural information processing systems*, 32.

Artem Shelmanov, Evgenii Tsymbalov, Dmitri Puzyrev, Kirill Fedyanin, Alexander Panchenko, and Maxim Panov. 2021. How certain is your transformer? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1833–1840.

Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer.

Dustin Tran, Jeremiah Liu, Michael W Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, et al. 2022. Plex: Towards reliability using pretrained large model extensions. *arXiv preprint arXiv:2207.07411*.

Dennis Ulmer. 2024. On uncertainty in natural language processing. *arXiv preprint arXiv:2410.03446*.

Liam Van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. *arXiv preprint arXiv:2210.13210*.

Vladimir Vapnik. 1991. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4.

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.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Akim Tsvigun, Daniil Vasilev, Rui Xing, Abdelrahman Boda Sadallah, Lyudmila Rvanova, Sergey Petrakov, Alexander Panchenko, et al. 2024. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *arXiv preprint arXiv:2406.15627*.

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, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. 2023. Efficient out-of-domain detection for sequence to sequence models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1430–1454.

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

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2020b. Minilmv2: Multihead self-attention relation distillation for compressing pretrained transformers. *arXiv preprint arXiv:2012.15828*.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696.

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*.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2024. The art of refusal: A survey of abstention in large language models. *arXiv preprint arXiv:2407.18418*.

Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

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.

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.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. In *International Conference on Learning Representations*.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019. Mitigating uncertainty in document classification. *arXiv preprint arXiv:1907.07590*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*.

A Implementation Details

We fine-tune the models end-to-end using the AdamW optimizer (Loshchilov, 2017), with a learning rate of 3e-5 and a batch size of 16. Training is conducted in a mixed-precision setting (fp16) to reduce memory usage. Models are trained for up to 20 epochs, with early stopping applied if validation performance does not improve for 3 consecutive epochs. Each experiment is repeated five times using different random seeds, and the mean results are reported. We utilize the FairLex datasets available on HuggingFace[2]. The λ parameter for Adversarial Removal, DeepCORAL, IRM, and V-REx is uniformly set to 0.5 across all datasets. We use mini-sized BERT models similar to those in FairLex, comprising 6 Transformer blocks, 384 hidden units, and 12 attention heads. Details regarding the HuggingFace model links and their corresponding datasets are provided in Table 2.

Dataset	Model
Domain-specific Pre-trained	
ECHR	`coastalcph/fairlex-ecthr-minilm`
SCOTUS	`coastalcph/fairlex-scotus-minilm`
FSCS	`coastalcph/fairlex-fscs-minilm`
CAIL	`coastalcph/fairlex-cail-minilm`
General Pre-trained	
ECHR	`nreimers/MiniLMv2-L6-H384-distilled-from-RoBERTa-Large`
SCOTUS	`nreimers/MiniLMv2-L6-H384-distilled-from-RoBERTa-Large`
FSCS	`nreimers/mMiniLMv2-L6-H384-distilled-from-XLMR-Large`
CAIL	`nreimers/mMiniLMv2-L6-H384-distilled-from-XLMR-Large`

Table 2: Datasets and their respective domain-specific and general purpose and pre-trained models.

B Results

Tab. 3, 4 presents the average macro-F1 (mF1), reliability measures (AURCC, RF) across groups and group disparity metrics (calculated for mF1, AURCC, and RF) across four datasets, under specific attribute, using domain-specific pre-trained model initializations. Higher mF1 indicates better overall performance, while lower AuRCC and Rf reflect greater reliability. Lower group disparity scores (D_{mF1}, D_{AuR}, D_{Rf}) suggest more equitable performance across groups under the evaluated attribute.

[2] https://huggingface.co/datasets/coastalcph/fairlex

ECHR																		
	Applicant Age						Applicant Gender						Defendant State					
	mF1	D_{mF1}	AuR	D_{AuR}	Rf	D_{Rf}	mF1	D_{mF1}	AuR	D_{AuR}	Rf	D_{Rf}	mF1	D_{mF1}	AuR	D_{AuR}	Rf	D_{Rf}
ERM	50.93	3.45	3.33	2.59	18.54	7.06	50.50	1.55	4.47	3.06	11.45	1.06	50.93	3.25	7.66	5.78	18.95	8.62
ERM+GS	**55.06**	3.34	**3.13**	**2.42**	17.92	7.03	51.44	**0.26**	4.20	2.85	12.21	1.50	51.44	**3.14**	**7.23**	**5.48**	**18.43**	8.38
AdvRem	51.44	6.41	3.71	2.98	21.23	6.22	**53.68**	0.78	4.76	3.35	13.90	**0.03**	**56.44**	5.97	8.66	5.95	19.93	**7.67**
DeepCoral	49.26	**2.92**	4.47	3.24	22.81	9.40	50.60	1.66	5.32	3.64	**11.32**	1.82	54.18	5.02	10.07	6.00	22.28	9.89
GDRO	50.70	3.65	3.25	2.71	**17.14**	5.76	51.42	1.65	**4.15**	**2.80**	14.73	3.48	51.76	4.22	8.41	6.32	19.07	8.34
IRM	51.84	4.12	4.31	3.11	19.35	7.03	51.54	2.42	5.07	3.23	15.54	3.90	51.88	3.62	8.99	6.85	18.98	7.81
V-REx	50.80	4.10	3.81	2.86	17.72	**5.68**	51.16	0.42	4.68	2.99	13.21	2.06	52.40	3.69	8.65	6.58	19.19	8.01
FSCS																		
	Court Region						Decision Language						Legal Area					
	mF1	D_{mF1}	AuR	D_{AuR}	Rf	D_{Rf}	mF1	D_{mF1}	AuR	D_{AuR}	Rf	D_{Rf}	mF1	D_{mF1}	AuR	D_{AuR}	Rf	D_{Rf}
ERM	**69.63**	3.59	**138.78**	**99.09**	**22.19**	4.82	**69.63**	3.02	**390.95**	256.17	**21.45**	1.66	**69.63**	9.28	**341.05**	**231.38**	**23.41**	5.81
ERM+GS	68.00	**2.94**	161.21	116.18	22.25	4.47	67.62	3.93	521.37	307.39	25.96	1.95	68.42	9.09	429.01	281.60	26.76	6.28
AdvRem	66.54	3.10	145.77	102.32	23.79	4.69	69.14	4.86	502.10	296.06	25.12	3.69	66.68	12.71	429.11	258.57	26.27	7.40
DeepCoral	68.42	3.51	165.77	112.68	24.46	4.76	66.60	3.51	459.97	268.12	24.29	2.34	67.38	**8.06**	415.16	268.64	25.39	6.02
GDRO	67.65	3.47	164.63	117.82	23.44	**3.94**	68.78	**1.73**	400.49	**251.48**	23.63	**1.44**	66.60	9.87	437.45	318.21	38.29	**5.66**
IRM	68.58	3.17	161.93	109.49	23.67	4.95	67.78	4.24	499.88	290.58	26.21	2.81	68.26	9.72	474.86	315.72	27.68	6.25
V-REx	68.18	3.38	161.03	114.97	23.18	5.20	67.23	3.99	488.96	284.05	25.87	2.69	68.60	9.70	392.41	254.61	24.90	5.91

Table 3: Results of bias mitigation algorithms on domain-specific pre-trained models applied to each attribute in ECHR and FSCS dataset. We report average of macro-F1 performance across groups (mF1), Reliability metrics (AuRCC (AuR), Rf) and Group Disparities in performance (D_{mF1}), AuRCC (D_{AuR}) and Refinement (D_{Rf}). For mF1 (↑), higher is better. For the rest, AuR (↓), Rf (↓), D_{mF1} (↓), D_{AuR} (↓) and D_{Rf} (↓), lower scores are better.

SCOTUS												
	Decision Direction						Respondent Type					
	mF1	D_{mF1}	AuR	D_{AuR}	Rf	D_{Rf}	mF1	D_{mF1}	AuR	D_{AuR}	Rf	D_{Rf}
ERM	83.00	1.25	**24.34**	0.36	18.82	1.27	**83.00**	4.03	**11.22**	**7.73**	21.86	2.48
ERM+GS	81.02	1.50	24.98	0.44	17.52	**0.83**	80.42	3.13	13.13	8.55	22.25	3.72
AdvRem	79.93	1.30	27.53	1.24	17.90	1.11	78.82	**2.21**	14.13	8.93	25.83	7.73
DeepCoral	81.14	1.93	26.61	1.61	18.17	1.54	80.10	3.17	13.73	9.96	**19.55**	**1.71**
GDRO	81.14	1.95	24.91	**0.16**	**17.24**	0.99	80.82	3.84	12.24	8.10	24.05	6.40
IRM	**83.10**	**1.21**	26.33	0.52	19.19	0.84	78.70	2.61	15.47	10.82	21.69	1.88
V-REx	81.94	1.66	25.59	1.41	19.16	1.95	81.30	3.37	12.93	9.12	22.38	2.80
CAIL												
	Court Region						Defendant Gender					
	mF1	D_{mF1}	AuR	D_{AuR}	Rf	D_{Rf}	mF1	D_{mF1}	AuR	D_{AuR}	Rf	D_{Rf}
ERM	59.23	3.15	**303.64**	**86.59**	**31.53**	**3.80**	**59.23**	0.23	**1019.07**	**867.83**	**30.69**	0.67
ERM+GS	58.62	2.90	323.64	90.85	31.79	4.20	55.32	0.86	1263.08	1068.08	33.98	0.29
AdvRem	58.33	2.72	310.74	92.85	32.31	4.49	53.98	**0.14**	1313.26	1104.21	34.92	0.19
DeepCoral	**59.53**	3.11	334.52	92.45	33.01	4.36	58.10	0.27	1198.03	1012.73	34.24	0.58
GDRO	58.27	**2.42**	317.58	89.30	31.84	4.17	58.23	0.93	1098.87	936.82	32.36	0.09
IRM	57.45	3.01	337.29	94.04	32.55	3.87	54.10	0.43	1276.12	1074.44	34.54	0.73
V-REx	58.26	3.08	319.64	92.34	31.76	4.54	56.84	0.44	1186.63	1002.14	33.31	**0.11**

Table 4: Results of various bias mitigation algorithms on domain-specific pre-trained models applied to each attribute in the SCOTUS and CAIL datasets. The notations are consistent with those in Table 3.