Does Fairness Survive Model Adaptation in Legal Judgment Prediction?

Sarah Kamoun, Parth Patel, Faezeh Ensan
Department of Electrical, Computer, and Biomedical Engineering
Toronto Metropolitan University
Toronto, Canada
{skamoun, parth5.patel, fensan}@torontomu.ca

Abstract

Large language models are increasingly used for Legal Judgment Prediction (LJP), raising concerns about demographic bias inherited from pre-training corpora and historical legal records. We investigate how fairness robustness changes under model adaptation in LJP through a controlled counterfactual evaluation across gender, ethnicity, and intersectional demographic groups. Using 23 demographic variants over encoder-based models and generative LLMs, we evaluate demographic sensitivity, counterfactual instability, and bias propagation under zero-shot, few-shot, and fine-tuned settings. Experiments show that model adaptation substantially affects fairness. Fine-tuning often improves predictive performance while simultaneously increasing demographic sensitivity and propagating biased supervision signals into downstream judicial predictions, whereas few-shot prompting generally produces smaller propagation effects. We further show that legal-domain specialization does not reliably guarantee fairness: legal-domain encoders span both highly robust and highly sensitive behavior after fine-tuning, and a legal-domain LLM is highly sensitive under few-shot prompting yet comparatively robust after fine-tuning under counterfactual perturbation.

1 Introduction

Legal Judgment Prediction (LJP) seeks to automatically infer judicial outcomes, such as charges, applicable statutes, or sentencing decisions, from legal texts like case fact descriptions, plaintiff complaints, and court summaries. LJP has become a central task in legal NLP due to its potential to support judicial decision-making and improve access to legal services. Recent advances in large language models (LLMs) have improved the performance of these systems, enabling them to better capture the reasoning patterns underlying legal decisions through large-scale pre-training (Chalkidis et al., 2020) and adaptation to legal tasks (Niklaus et al., 2025; Li et al., 2025; T.Y.S.S et al., 2024).

Despite the gains in predictive performance, LLM-based legal systems remain vulnerable to biases inherited from large-scale training corpora and historical judicial records. Prior work has shown that language models can learn, reproduce, and amplify demographic stereotypes and unequal treatment patterns across social groups (Gallegos et al., 2024; Delobelle et al., 2022). In legal contexts, such biases are particularly concerning because judicial predictions directly affect rights, sentencing outcomes, and public trust (Washington, 2018). If identical case facts lead to different predictions solely due to changes in demographic attributes such as gender or ethnicity, the system is relying on factors that should be legally irrelevant. This raises concerns that LJP systems are potentially reinforcing biases embedded not only in historical legal records but also in the broader world knowledge and social stereotypes encoded within LLMs during pre-training.

Existing research on fairness in legal judgment prediction has largely focused on developing benchmarks, evaluation protocols, and datasets for measuring judicial bias and inconsistency (Zhang et al., 2024; Wang et al., 2021; Zhang et al., 2025). While these efforts establish important foundations for measuring fairness, questions regarding the origins and behavior of bias in LJP systems remain underexplored. In particular, limited attention has been given to how knowledge acquired during pre-training interacts with external legal knowledge during inference, how robust LJP models remain under few-shot and fine-tuning settings, whether biased fine-tuning data amplifies demographic disparities, and whether legal LLMs are more robust or fair than general-purpose models. These challenges are especially important in common law litigation settings, where demographic information is often implicitly embedded within the case nar-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5653–5667
November 6-10, 2026 ©2026 Association for Computational Linguistics

ratives rather than expressed in explicit structured attributes (Chalkidis et al., 2022).

In this work, we present a comprehensive study investigating fairness and demographic bias in LLM-based legal judgment prediction across both legal-domain-specific language models and general-purpose LLMs. Our analysis considers multiple demographic settings, including variations in gender, ethnicity, and intersectional cases where individuals belong to multiple demographic groups simultaneously (Lalor et al., 2022). We further compare the relationship between predictive performance and fairness under zero-shot, few-shot, and fine-tuned settings to better understand how prompting strategies and task adaptation influence model behavior. In addition, we examine the robustness of these models when exposed to deliberately biased supervision signals, analyzing whether bias propagates into judicial predictions or whether models can recover the underlying legal reasoning despite the presence of biased information.

We investigate the following research questions:
RQ1: To what extent do LLM-based LJP systems exhibit demographic bias across variations in gender, ethnicity, and intersectional identity groups? How do legal-specific models differ from general-purpose LLMs with respect to fairness, and which demographic groups are most or least affected by prediction disparities?
RQ2: How do zero-shot prompting, few-shot prompting, and fine-tuning influence the fairness and robustness of LJP models? More specifically, how are gains in predictive performance balanced against potential increases in fairness vulnerabilities across different adaptation settings?
RQ3: When fine-tuning data or few-shot demonstrations contain biased judicial outcomes, to what extent does this bias propagate into model predictions? Which demographic groups are most vulnerable to such propagation effects, and how robust are different models to biased supervision signals?

Our contributions can be summarized as follows:

- We conduct a controlled counterfactual stress test of 9 LJP models across 23 demographic variants, evaluating counterfactual sensitivity and group fairness under gender, ethnicity, and intersectional settings.
- We show that model choice affects both the magnitude and direction of demographic sensitivity across both encoders and generative LLMs.
- We compare zero-shot, few-shot, and fine-tuned adaptation settings, demonstrating that legal-domain pre-training does not guarantee demographic robustness.
- We analyze the tradeoff between predictive accuracy and demographic robustness, showing that fine-tuning often improves effectiveness while increasing fairness disparities and counterfactual instability.
- We conduct bias injection experiments in both fine-tuning and few-shot prompting settings, showing that biased supervision signals propagate into downstream judicial predictions with varying strength across model families and adaptation strategies.

2 Related Work

Legal Judgment Prediction. Legal judgment prediction frames case outcome forecasting as a text classification problem. The field has moved from feature-based classifiers through BERT-era encoders (Chalkidis et al., 2019, 2020) to recent LLM and reasoning-based approaches (Shi et al., 2025; Zhang et al., 2025; Shukla et al., 2026), with surveys covering its datasets, metrics, and models (Cui et al., 2023). Its empirical landscape is dominated by a handful of jurisdictions: Chinese criminal cases (Xiao et al., 2018), European Court of Human Rights filings (Chalkidis et al., 2019), Swiss court decisions (Niklaus et al., 2021), Indian Court cases (Malik et al., 2021; Nigam et al., 2024), and UK employment tribunals (Xie et al., 2024), with US civil-litigation resources such as ClassActionPrediction (Semo et al., 2022) comparatively rare. The field remains accuracy-dominated and skews toward non-US jurisdictions; fairness in these systems, particularly in US civil settings, is not the question most LJP work asks.
Fairness and Bias in Legal AI. Fairness auditing is well established in NLP broadly (Gallegos et al., 2024; Delobelle et al., 2022), but remains comparatively sparse in the legal domain. Existing work includes the multilingual fairness benchmark FairLex (Chalkidis et al., 2022), group-reliability disparities on ECtHR (T.Y.S.S and Chowdhury, 2025), judgment consistency analysis for fairness (Wang et al., 2021), and recent multi-model LLM audits such as the JudiFair benchmark (Hu et al., 2025) and the JustEva toolkit (Xue et al., 2025). This literature concentrates on Chinese criminal, European, and Indian jurisdictions, and its fairness subset inherits that same skew. Existing multi-model LLM audits

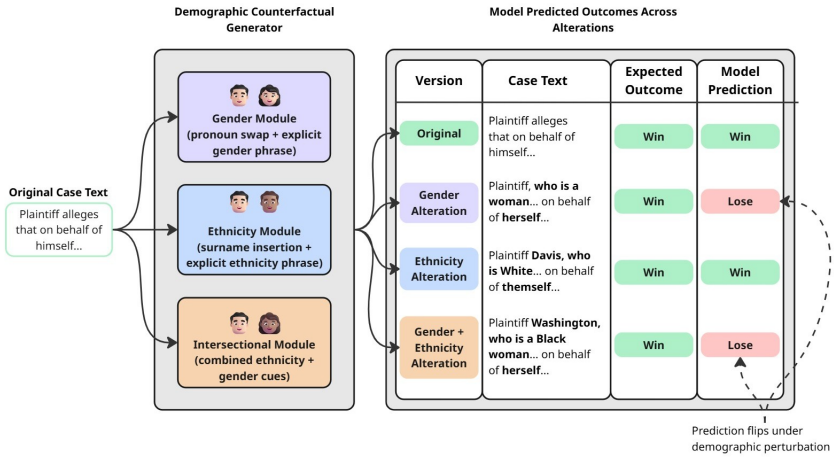


Figure 1: Counterfactual demographic perturbation pipeline for evaluating prediction sensitivity under gender, ethnicity, and intersectional alterations.

compare general-purpose LLMs without systematically evaluating them under model adaptation or contrasting them against legal-specific models as explored in our study.

Counterfactual and Injection-Based Bias Evaluation. Our methodology draws on counterfactual fairness (Kusner et al., 2017) and on NLP audits that alter or remove sensitive attributes while holding the remaining text fixed. Single-attribute counterfactual injection is precedented within LJP: LJPCheck (Zhang et al., 2024) applies functional tests that flip or add name, gender, and ethnicity cues, and recent LLM audits perform counterfactual demographic swaps over case facts (Hu et al., 2025; Xue et al., 2025). Prior work swaps explicit demographic labels, treats attributes one at a time, and largely operates outside common law jurisdictions. We instead layer ethnic surnames, pronoun substitution, and an explicit demographic clause together across gender, ethnicity, and their intersection in common law litigation, additionally analyzing the direction of each flip to characterize the type of bias a model exhibits, not just its magnitude.

Bias Propagation and Amplification. Our propagation study builds on the bias amplification literature, which formalizes how models can magnify disparities present in their training data (Zhao et al., 2017) and quantifies this directionally (Wang and Russakovsky, 2021). In-context demonstrations and fine-tuning data are both known to inject bias into downstream predictions in general NLP; in LJP specifically, bias propagation has received limited attention. Our contribution is a controlled bias-injection design: verdicts are deterministically coupled to a demographic attribute together with a mirror condition (positive coupling in one training set, reversed coupling in a second), and we measure the resulting shift in predicted win probability between favored and contrasting variants, yielding a signed, bounded measure of how strongly the injected coupling transfers to unseen cases.

3 Methodology

In this section, we describe the experimental framework used to investigate demographic bias in LLM-based legal judgment prediction across gender, ethnicity, and intersectional identity groups.

3.1 Dataset

We use ClassActionPrediction (Semo et al., 2022), a corpus of U.S. civil class-action cases paired with binary win/lose verdicts. Following the original train-validation-test split (70%/15%/15%), our evaluation is conducted on a test set of 450 cases. Unlike many prior LJP datasets derived from civil law systems or based on court-written fact summaries (Cui et al., 2023), ClassActionPrediction uses plaintiff complaints as input, capturing realistic pre-judgment common law litigation narratives in which demographic information is conveyed both explicitly and implicitly through names, contextual details, and case descriptions. This makes the dataset suitable for evaluating fairness and robustness in LLM-based legal judgment prediction.

3.2 Counterfactual Variant Construction

We construct demographic counterfactual variants using a rule-based pipeline that identifies the primary individual referenced in each case and applies demographic modifications only when the entity

refers to a person rather than a company or organization. Each class action is filed by a named representative plaintiff, and only this plaintiff is perturbed (Appendix F). Variants are generated through three channels. *Gender* variants modify pronouns and introduce an explicit demographic phrase (e.g., "who is a woman/man"). *Ethnicity* variants set the plaintiff's surname to an ethnically associated one and add an explicit demographic phrase (e.g., "who is [ethnicity]"). For this purpose, we use ethnicity-associated surnames derived from prior demographic surname statistics (United States Census Bureau, 2016). For example, surnames such as "Washington" and "Hu" are strongly associated with particular demographic groups in population records. *Intersectional* variants combine both gender and ethnicity cues (e.g., "who is a [ethnicity] woman/man"). Overall, the framework generates two gender variants (Man and Woman), seven ethnicity variants, and fourteen intersectional variants, resulting in 23 demographic variants per case. The ethnicity groups considered in this work are Black, White, East Asian, Hispanic, Middle Eastern and North African (MENA), Slavic, and South Asian. This construction strategy is informed by prior work on counterfactual fairness (Kusner et al., 2017) and demographic perturbation methods in NLP (Zhang et al., 2024). Figure 1 illustrates the counterfactual data generation process.

3.3 Evaluated Models

We evaluate nine models spanning both legal-domain-specific and general-purpose language model families. The evaluated models include encoder-based architectures, namely BERT, Legal-BERT (Chalkidis et al., 2020), CaseLawBERT, and LegalRoBERTa, as well as open generative LLMs including Llama-3.1-8B, Llama-3.1-8B-Instruct, Qwen-2.5-7B, Qwen-2.5-7B-Instruct and Saul-7B-Instruct-v1 (Colombo et al., 2024). Table 4 in Appendix A summarizes the evaluated models.

3.4 Evaluation Protocol

For the primary fairness evaluation, we evaluate models under three settings: zero-shot, few-shot, and fine-tuned. A key aspect of our experimental design is that all training and prompting procedures use the original case text, while demographic variants are introduced only at inference time. This allows changes in model predictions to be attributed to demographic perturbations rather than differences in training distribution. For few-shot prompting, we use $k = 2$ labelled demonstrations. All three evaluation settings follow a shared prompt structure, shown in Figure 5 in Appendix B. For the fine-tuned setting, models are fine-tuned using the original training split of the dataset.

To investigate whether demographic biases introduced during fine-tuning propagate into downstream predictions, we fine-tune models on systematically biased training corpora in which verdict labels are explicitly correlated with demographic attributes. Specifically, we construct training distributions where positive and negative verdicts are deterministically associated with specific demographic variants. In the *Female Bias* setting, all training instances involving female demographic variants are positive verdict cases, while male variants are negative verdict cases. The *Male Bias* setting reverses this by pairing male variants with positive verdicts and female variants with negative verdicts. Similarly, the *Black Bias* and *White Bias* settings associate positive verdicts with Black and White demographic variants, respectively.

During evaluation, the model that was fine-tuned on biased training data is used to predict the winning probability of three versions of each test case. The first is anonymized, where all demographic references are removed. The second version corresponds to the demographic group favored during training. For example, under the *Female Bias* setting, the plaintiff is represented as female. The third version represents the same case using the contrasting demographic variant, such as a male plaintiff. Inspired by the directional bias concept proposed in Wang and Russakovsky (2021), we measure how biases introduced during training propagate into downstream predictions by comparing the difference in predicted winning probabilities between the favored demographic variant and the anonymized version, as well as between the non favored demographic variant and the anonymized version.

For the few-shot setting, we follow the same methodology, but instead of fine-tuning the model, we condition it using $K = 4$ biased few-shot exemplars provided in the prompt. For example, in the *Female Bias* configuration, we construct few-shot examples consisting of two winning cases involving female plaintiffs and two losing cases involving male plaintiffs. During evaluation, we again generate anonymized, favored demographic, and contrasting demographic versions of each test case and compute the predicted probability of a case verdict under the biased few-shot context. To ensure the

validity of counterfactual pairs, we identified and excluded cases where the plaintiff's demographic attribute is legally relevant. More details on data pre-processing are presented in Appendix E.

3.5 Evaluation metrics

We report Accuracy and Macro-F1 to evaluate predictive performance, together with fairness evaluation measures. Following prior work on counterfactual fairness (Kusner et al., 2017), we measure the *Counterfactual Flip Rate* (CFR), which quantifies how often a model changes its prediction after demographic information is modified while the remaining case facts are held fixed. Higher CFR values indicate greater sensitivity to demographic perturbations. We additionally evaluate group-level fairness using *Demographic Parity Gap* (DP-Gap) (Dwork et al., 2012), defined as the difference between the maximum and minimum positive prediction rates across demographic groups. Lower DP-Gap values indicate more balanced prediction behavior across the different demographic identities. We report DP-Gap separately for gender, ethnicity, and intersectional demographic groups.

Finally, to evaluate whether demographic associations introduced during fine-tuning propagate into downstream predictions, we adopt the *Propagation Score* τ, based on the probabilities that a model assigns to a *win* verdict for a given case. In the fine-tuning setting, this probability is produced by a model trained on biased data favoring the demographic variant. In the few-shot setting, the probability is produced by a model conditioned on biased exemplars favoring the same demographic variant. A positive propagation score τ indicates that the injected demographic bias transfers into downstream predictions in the same direction as the supervision signal, leading to higher predicted winning probabilities for the favored demographic group. A negative value indicates propagation in the opposite direction, where the model favors the demographic group disfavored during training. Values of τ close to zero indicate limited or no measurable propagation of demographic bias into downstream legal predictions. More details about the evaluation metrics have been presented in Appendix C.

4 Results

4.1 Existence of Demographic Bias

Table 1 presents the predictive performance and fairness evaluation results for models fine-tuned on the original training split of the dataset. Predictive effectiveness is evaluated using Accuracy and Macro-F1, while fairness is assessed using demographic parity gap (DP Gap) and counterfactual flip rate (AVG CFR). Gender fairness measures are averaged across the two gender groups, Male and Female, whereas ethnicity fairness measures are averaged across seven ethnicity groups, namely Black, East Asian, Hispanic, MENA, Slavic, South Asian, and White. Intersectional fairness scores are averaged across all 14 gender-ethnicity combinations. The models exhibit distinct fairness and predictive performance characteristics across encoder-based legal language models and generative LLMs, with noticeable variation in how demographic robustness changes across architectures, instruction tuning settings, and model scales.

Among the encoder-based legal models, *Legal-RoBERTa* achieves the strongest balance between predictive effectiveness and fairness robustness, attaining high Accuracy (0.6289) and Macro-F1 (0.6194) while maintaining consistently low DP Gap and AVG CFR values across all demographic settings (e.g. 0.0302 DP-Gap and 0.0221 AVG CFR for intersectional identities). In contrast, *CaseLawBERT* exhibits substantially larger demographic disparities (e.g. 0.1465 DP-Gap and 0.1357 AVG CFR) despite comparable predictive performance, while *LegalBERT* occupies an intermediate position in balancing performance and fairness. This all suggests that legal-domain specialization alone does not guarantee fairness robustness.

Among generative models, model scale and instruction tuning appear to influence both predictive quality and fairness outcomes. The *Llama-3.1-8B* model achieves the best overall predictive performance (Table 1) while maintaining moderate fairness disparities. The instruction tuned variant, *Llama-3.1-8B-Instruct*, slightly reduces predictive performance while increasing fairness disparities.

Gender bias. Figure 2 illustrates prediction disparities across genders by comparing the probability of predicting a winning verdict for anonymized, male, and female variants of the same case, together with the corresponding *win_to_lose rate*, which measures how often a case predicted as a win in its anonymized form changes to a loss after demographic substitution. Across most models, gender perturbations produce relatively small changes (below $\Delta = 0.04$), yet distinct gender preference patterns emerge. Among legal encoders, *CaseLawBERT* slightly favors female vari-

Table 1: Performance and fairness evaluation results for models fine-tuned on the original training split of the dataset.

Model	Performance ↑		DP Gap ↓			AVG CFR ↓		
	Accuracy	Macro-F1	Gender	Ethnicity	Intersectional	Gender	Ethnicity	Intersectional
BERT	0.6222	0.6144	0.0302	0.0439	0.0628	0.0616	0.0749	0.0982
LegalBERT	0.6000	0.5989	0.0209	0.0785	0.0814	0.0314	0.0426	0.0547
CaseLawBERT	0.6000	0.5999	0.0442	0.1409	0.1465	0.1035	0.1359	0.1357
LegalRoBERTa	0.6289	0.6194	0.0349	0.0300	0.0302	0.0116	0.0191	0.0221
Llama-3.1-8B-Instruct	0.6222	0.6217	0.0884	0.0878	0.0907	0.0605	0.0851	0.0912
Qwen-2.5-7B-Instruct‡	0.5178	0.3717	0.0070	0.0139	0.0279	0.0105	0.0112	0.0150
Llama-3.1-8B	0.6378	0.6374	0.0465	0.0600	0.0628	0.0465	0.0429	0.0447
Qwen-2.5-7B‡	0.5067	0.4152	0.1558	0.1570	0.1558	0.1163	0.1092	0.1199
Saul-7B-Instruct-v1	0.6336	0.6333	0.0000	0.0139	0.0208	0.0093	0.0106	0.0364

‡LoRA fine-tuning on the original training split had no measurable effect on either Qwen model: Accuracy and Macro-F1 are identical to zero-shot (Table 2).

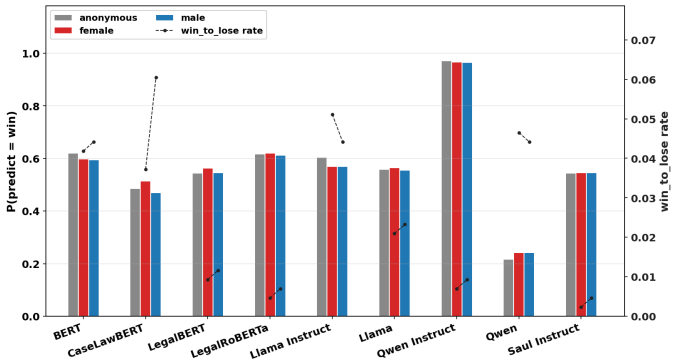


Figure 2: Gender based counterfactual sensitivity across legal encoders and generative LLMs.

ants and exhibits the highest *win_to_lose rate*, indicating greater gender sensitivity. In contrast, *LegalBERT* slightly favors male variants, while *LegalRoBERTa* produces nearly identical probabilities across genders. Among generative LLMs, gender effects are weaker. *Llama-3.1-8B* and *Llama-3.1-8B-Instruct* show mild female favoring tendencies, with the instruction tuned variant exhibiting slightly higher *win_to_lose rates*, suggesting increased demographic sensitivity after instruction tuning.

Ethnicity bias. Figure 3 illustrates prediction disparities across seven ethnicity groups. Compared to gender, ethnicity perturbations produce substantially larger variations across models. Among encoder-based legal models, *CaseLawBERT* exhibits the strongest ethnicity sensitivity and the largest variation across demographic groups. *LegalBERT* also demonstrates noticeable ethnicity preferences, slightly favoring Black, Hispanic, and White variants over East Asian and Slavic groups. In contrast, *LegalRoBERTa* produces considerably more stable predictions across ethnicities, with only minor variations between groups and near zero *win_to_lose rate* values. Among generative LLMs, ethnicity related disparities are generally smaller but still observable. *Llama-3.1-8B-Instruct* demonstrates moderate variation across groups, with lower predictions for Black and Hispanic variants and comparatively higher probabilities for MENA and South Asian groups. The base *Llama-3.1-8B* model exhibits more stable behavior overall, with relatively uniform predictions across ethnicities and low *win_to_lose rate* values.

Intersectional bias. Figure 4 extends the analysis to intersectional demographic groups by jointly considering gender and ethnicity variations. Across nearly all models, certain groups consistently benefit from higher predicted winning probabilities, while others experience systematically lower predictions and higher *win_to_lose* rates. Among legal encoders, *CaseLawBERT* again exhibits the strongest demographic instability. In particular, Hispanic female and Black female variants often receive higher winning probabilities than their male counterparts, while MENA female and Slavic female groups experience some of the lowest predictions among all intersectional combinations. Among generative LLMs *Llama-3.1-8B-Instruct* exhibits noticeable variation across demographic combinations, with MENA male and South Asian

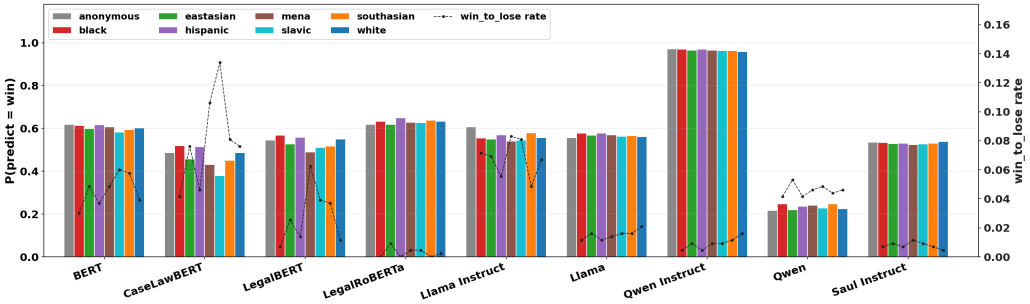


Figure 3: Ethnicity based counterfactual sensitivity across legal encoders and generative LLMs.

Note: Bars show P(verdict = win) per demographic variant on the left y-axis; the dotted line (right y-axis) is the win-to-lose rate, which is the fraction of cases predicted as win under the anonymized variant that flip to lose under a demographic variant.

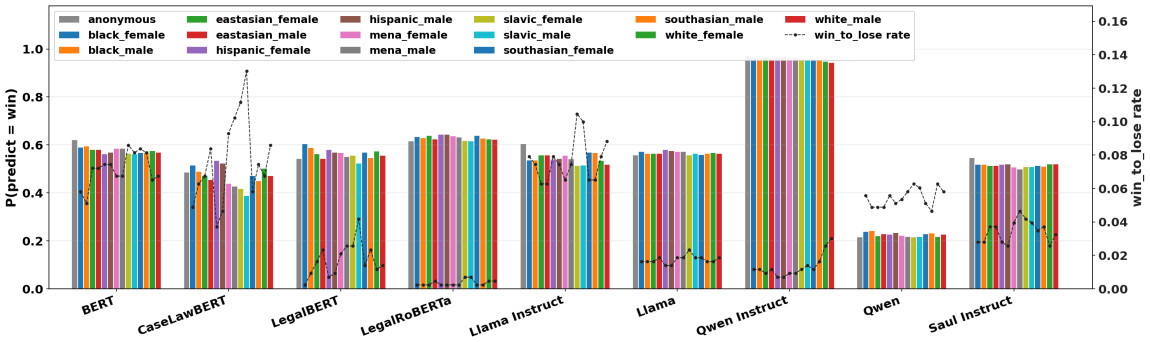


Figure 4: Intersectional based counterfactual sensitivity across legal encoders and generative LLMs.

male variants often receiving slightly higher winning probabilities than Black or White male groups.

4.2 Variation Across Adaptation Settings

Table 2 compares zero-shot, few-shot, and fine-tuned variants of encoder-based models and generative LLMs in terms of predictive performance and demographic robustness. Across most model families, fine-tuning substantially improves predictive effectiveness, often producing large gains in both Accuracy and Macro-F1, but these improvements are frequently accompanied by increased demographic sensitivity. For example, fine-tuning *BERT*, *CaseLawBERT*, and *Llama-3.1-8B-Instruct* significantly improves classification performance while simultaneously increasing fairness disparities across gender, ethnicity, and intersectional settings. Few-shot prompting generally introduces smaller fairness shifts than fine-tuning, although its effects vary considerably across models. For instance, few-shot prompting slightly increases demographic sensitivity in *Qwen-2.5-7B-Instruct* and *Llama-3.1-8B-Instruct*, while improving fairness robustness, at some accuracy cost, for the base *Qwen-2.5-7B* model. Near-zero CFR in zero-shot settings reflects prediction collapse rather than robustness, and achieve substantially weaker predictive performance than their fine-tuned counterparts.

4.3 Bias Propagation

Table 3 reports the propagation score τ for demographic biases injected through fine-tuning and few-shot prompting across gender and ethnicity settings. Higher τ values indicate greater transfer of demographic associations into downstream legal predictions and therefore lower robustness to biased supervision, while values near zero indicate stronger resistance to demographic bias transfer. Encoder-based models consistently exhibit high propagation scores under fine-tuning, with gender and ethnicity propagation around 0.4. Among generative LLMs, all fine-tuned models except *Qwen-2.5-7B-Instruct* exhibit high propagation effects, with male bias injection producing the highest observed propagation (0.467356 for *Llama-3.1-8B*). In contrast, few-shot prompting produces substantially smaller propagation scores, typically below 0.02.

4.4 Empirical Findings

Intersectional Bias is Additive. The results consistently show that demographic sensitivity becomes substantially stronger when gender and eth-

Table 2: Comparison of base, few-shot, and fine-tuned model variants across performance and fairness metrics. Values in parentheses denote the change in CFR relative to the corresponding base or zero-shot configuration.

Model	Setting	Performance		AVG CFR (Δ from Base)		
		Accuracy	Macro-F1	Gender	Ethnicity	Intersectional
BERT	Base	0.5178	0.3751	0.0035	0.0066	0.0078
	Fine-tuned	0.6222	0.6144	0.0616 (+0.0581)	0.0749 (+0.0683)	0.0982 (+0.0904)
LegalBERT	Base	0.5222	0.4750	0.1186	0.1201	0.1618
	Fine-tuned	0.6000	0.5989	0.0314 (-0.0872)	0.0426 (-0.0775)	0.0547 (-0.1071)
CaseLawBERT	Base	0.4867	0.3348	0.0116	0.0030	0.0068
	Fine-tuned	0.6000	0.5999	0.1035 (+0.0919)	0.1359 (+0.1329)	0.1357 (+0.1289)
LegalRoBERTa	Base	0.5089	0.3373	0.0000	0.0000	0.0000
	Fine-tuned	0.6289	0.6194	0.0116 (+0.0116)	0.0191 (+0.0191)	0.0221 (+0.0221)
Llama-3.1-8B-Instruct	Zero-shot	0.5089	0.3373	0.0000	0.0000	0.0000
	Few-shot	0.5156	0.3522	0.0023 (+0.0023)	0.0056 (+0.0056)	0.0060 (+0.0060)
	Fine-tuned	0.6222	0.6217	0.0605 (+0.0605)	0.0851 (+0.0851)	0.0912 (+0.0912)
Qwen-2.5-7B-Instruct	Zero-shot	0.5178	0.3717	0.0105	0.0115	0.0145
	Few-shot	0.5533	0.4648	0.0221 (+0.0116)	0.0340 (+0.0225)	0.0375 (+0.0230)
	Fine-tuned	0.5178	0.3717	0.0105 (+0.0000)	0.0112 (-0.0003)	0.0150 (+0.0005)
Llama-3.1-8B	Zero-shot	0.5089	0.3373	0.0000	0.0000	0.0000
	Few-shot	0.5089	0.3373	0.0000 (+0.0000)	0.0000 (+0.0000)	0.0000 (+0.0000)
	Fine-tuned	0.6378	0.6374	0.0465 (+0.0465)	0.0429 (+0.0429)	0.0447 (+0.0447)
Qwen-2.5-7B	Zero-shot	0.5067	0.4152	0.1140	0.1095	0.1201
	Few-shot	0.4600	0.4289	0.0895 (-0.0245)	0.0970 (-0.0125)	0.1043 (-0.0158)
	Fine-tuned	0.5067	0.4152	0.1163 (+0.0023)	0.1092 (-0.0003)	0.1199 (-0.0002)
Saul-7B-Instruct-v1	Zero-shot	0.4965	0.3318	0.0000	0.0000	0.0000
	Few-shot	0.5196	0.4403	0.0651 (+0.0651)	0.0617 (+0.0617)	0.0610 (+0.0610)
	Fine-tuned	0.6336	0.6333	0.0093 (+0.0093)	0.0106 (+0.0106)	0.0364 (+0.0364)

nicity are considered jointly rather than independently. Across nearly all fine-tuned models, intersectional perturbations produce higher CFR values than gender-only or ethnicity-only modifications (for example BERT's CFR is 0.0982 for intersectional vs. 0.0616 gender and 0.0749 ethnicity; LegalBERT: 0.0547 vs. 0.0314 and 0.0426), indicating that evaluations based on isolated demographic attributes underestimate the instability models may exhibit in realistic legal settings.

Demographic Parity and Counterfactual Stability Capture Different Bias Modes. The experiments further show that DP-Gap and CFR measure fundamentally different forms of demographic sensitivity. For instance, Saul-7B-Instruct-v1 has a gender DP-Gap of 0 but a CFR of 0.0093, and CaseLawBERT has a gender DP-Gap of 0.0442 but a CFR of 0.1035. While outcome distributions appear balanced at the group level, individual predictions remain sensitive to demographic modifications. This demonstrates that parity-based evaluations may overlook demographic dependence that directly affects judicial outcomes in otherwise equivalent cases.

Fine-Tuning Improves Accuracy but Often Amplifies Demographic Sensitivity. Table 2 reveals a consistent tradeoff between predictive performance and demographic robustness. Supervised fine-tuning generally improves legal judgment prediction accuracy by enabling models to better capture domain-specific reasoning patterns (for example: 0.5156 for few-shot Llama-3.1-8B-Instruct vs. 0.6222 fine-tuned). However, these gains are frequently accompanied by increased counterfactual sensitivity: gender CFR rises from 0.0023 to 0.0605 for the same Llama model when fine-tuned.

Legal-Domain Pre-training Provides No Guarantee for Improved Fairness. Among the encoders, fine-tuned BERT and the fine-tuned legal-domain encoders span the full range of counterfactual sensitivity across all three demographic axes (Table 1). For gender, average CFR ranges from 0.0116 for LegalRoBERTa-ft to 0.1035 for CaseLawBERT-ft, while the general-purpose BERT-ft baseline lies between them at 0.0616. Legal-domain specialization therefore does not consistently reduce counterfactual sensitivity.

Adaptation Setting Determines Whether Legal-Domain Specialization Helps. With two in-context demonstrations, Saul-7B-Instruct reaches an accuracy of only 0.5196 while its CFR rises to 0.0651 gender, 0.0617 ethnicity and 0.0610 intersectional, the largest few-shot increase of any generative model. Fine-tuning reverses this: accuracy

Table 3: Propagation score τ for gender and ethnicity bias injection across fine-tuned and few-shot model variants.

Model	Variant	Gender Propagation τ		Ethnicity Propagation τ	
		Female Bias	Male Bias	Black Bias	White Bias
BERT	Fine-tuned	0.435310	0.435663	0.402102	0.402097
LegalBERT	Fine-tuned	0.434638	0.436157	0.400455	0.401675
CaseLawBERT	Fine-tuned	0.435195	0.434809	0.401270	0.400739
LegalRoBERTa	Fine-tuned	0.435108	0.435231	0.401833	0.401984
Llama-3.1-8B-Instruct	Fine-tuned	0.443272	0.462979	0.441078	0.440036
	Few-shot	0.007767	0.002474	0.010894	0.002405
Qwen-2.5-7B-Instruct	Fine-tuned	0.003892	-0.003892	0.006266	-0.006266
	Few-shot	0.000007	0.003115	0.014986	0.005854
Llama-3.1-8B	Fine-tuned	0.452814	0.467356	0.441102	0.441085
	Few-shot	0.013459	0.010316	0.019102	0.012089
Qwen-2.5-7B	Fine-tuned	0.438215	0.299194	0.438572	0.417167
	Few-shot	0.010991	0.009980	0.022254	0.016431
Saul-7B-Instruct-v1	Fine-tuned	0.465969	0.465019	0.437798	0.437634

reaches 0.6336 and macro-F1 0.6333, second only to Llama-3.1-8B, while CFR falls to 0.0093 gender and 0.0106 ethnicity, the lowest of any fine-tuned generative model.

Low Counterfactual Flip Rates Do Not Guarantee Robustness to Biased Supervision. Among the fine-tuned encoder models, the gender CFR ranges from 0.0116 for LegalRoBERTa to 0.1035 for CaseLawBERT. However, under the propagation-injection protocol in Table 3, every fine-tuned encoder shifts its predicted probability of a win toward the injected demographic direction by a similar magnitude. Notably, LegalRoBERTa has the lowest audit-time CFR at 0.0116, yet its propagation shift is comparable to that of the other encoders. This indicates that stability under test-time demographic perturbations is different from resistance to demographic bias introduced through supervision. The same distinction appears for the generative models. Saul-7B-Instruct-zs has a CFR of 0 on every demographic axis, but this is not evidence of meaningful robustness. The model behaves as a degenerate constant predictor and achieves an accuracy of only 0.4965 and a macro-F1 of 0.3318; three zero-shot models produce identical predictions across the entire test set. By comparison, LegalRoBERTa-ft has a similarly low CFR of 0.0116 while achieving a substantially higher accuracy of 0.6289. Thus, a low CFR can reflect either prediction collapse, as in Saul-7B-Instruct-zs, or genuine counterfactual invariance, as more plausibly reflected by LegalRoBERTa-ft. The dissociation is clear for Saul-7B-Instruct-v1, which has the lowest CFR of fine-tuned generative models (0.0093 gender, 0.0106 ethnicity), yet its propagation scores under bias injection are 0.465969 and 0.437798 respectively, among the highest observed across all models. CFR alone is therefore not sufficient to establish robustness, and audit-time CFR should be reported alongside supervision measurements as in Table 3.

5 Conclusion

We presented a fairness evaluation of legal judgment prediction models across 23 counterfactual demographic variants in common-law civil class-action cases. The results show that demographic bias remains widespread across both legal-domain encoders and general-purpose LLMs, with intersectional demographic groups exhibiting the strongest instability. Importantly, models differ not only in the magnitude but also in the direction of demographic sensitivity, indicating that model substitution may redirect rather than eliminate demographic harm. Adaptation setting matters as much as model choice: fine-tuning improves accuracy while amplifying counterfactual sensitivity, and injected demographic couplings transfer into predictions far more strongly under fine-tuning than few-shot prompting. Audit-time counterfactual stability does not predict resistance to biased supervision, so both should be reported. For future work, we aim to evaluate whether the observed demographic disparities and counterfactual flip rates persist across a range of prompt variants, as prompt-level interactions with demographic cues represent an additional source of fairness risk that the current study does not capture. Possible bias mitigation directions are discussed in Appendix G.

6 Limitations

Deliberately Strong Demographic Signals. The explicit phrase is repeated at almost every plaintiff mention, maximizing the demographic signal. The design is a deliberate stress test, and the resulting disparities should be read as an upper bound.

Dataset Generalization. Our results are reported on a common law US dataset with plaintiff complaints as case text and binary win/lose labels as verdicts. As such, we do not generalize on datasets that represent other legal systems (e.g. civil law), languages (e.g. Chinese), jurisdictions (e.g. U.K.), or input types (e.g. court-authored summaries). The binary outcome also implies that disparities that appear only in degree, such as partial wins or settlements, are unobservable. As such, datasets such as ILDC (Malik et al., 2021) (India), CLC-UKET (UK)(Xie et al., 2024), LegalHK (Shi et al., 2025) (Hong Kong) or US Supreme Court data (Alali et al., 2021) are left for future work.

Possibility of Data Contamination. A limitation of evaluating recent large language models on the US Class Actions dataset is the possibility of data contamination. The dataset was released in 2022; verdicts for some cases in the corpus, or paraphrases of those verdicts, may have appeared in the models' pre-training data. However, average CFR and DP-Gap are computed using within-case comparisons between anonymized and demographically perturbed versions of the same legal case. Any memorized information would therefore be shared across the compared versions and would generally affect their predictions in the same direction, reducing its effect on the measured counterfactual difference. Nevertheless, contamination cannot be completely ruled out and remains a limitation of the evaluation.

Limited Gender and Ethnicity Categories. Our counterfactual variants cover a limited set of gender and ethnicity categories. In particular, the gender perturbations are restricted to binary man/woman variants, and the ethnicity perturbations cover seven broad groups. A more complete demographic audit would consider non-binary and gender-diverse identities, finer-grained ethnic and racial categories, and additional demographic dimensions such as age, religion, disability, and socioeconomic status.

Open-source Model Coverage. Our model evaluation is limited to open-source encoder models and open-weight generative LLMs. We do not evaluate closed proprietary systems, which are increasingly used in practical legal-AI deployments and may exhibit different fairness, robustness, and bias-propagation behavior. Future work should extend the same counterfactual evaluation protocol to closed-source commercial models when access and reproducibility constraints permit.

References

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

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

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

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and 1 others. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, 11:102050–102071.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

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.

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.

Yiran Hu, Zongyue Xue, Haitao Li, Siyuan Zheng, Qingjing Chen, Shaochun Wang, Xihan Zhang, Ning Zheng, Yun Liu, Qingyao Ai, and 1 others. 2025. Llms on trial: Evaluating judicial fairness for large language models. *arXiv preprint arXiv:2507.10852*.

Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4069–4079, Red Hook, NY, USA. Curran Associates Inc.

John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in nlp. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

Shangyuan Li, Shiman Zhao, Zhuoran Zhang, Zihao Fang, Wei Chen, and Tengjiao Wang. 2025. Basis is also explanation: Interpretable legal judgment reasoning prompted by multi-source knowledge. *Information Processing & Management*, 62(3):103996.

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

Shubham Kumar Nigam, Anurag Sharma, Danush Khanna, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2024. Legal judgment reimagined: PredEx and the rise of intelligent AI interpretation in Indian courts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4296–4315, Bangkok, Thailand. Association for Computational Linguistics.

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

Joel Niklaus, Lucia Zheng, Arya D McCarthy, Christopher Hahn, Brian M Rosen, Peter Henderson, Daniel E Ho, Garrett Honke, Percy Liang, and Christopher D Manning. 2025. Lawinstruct: A resource for studying language model adaptation to the legal domain. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 127–152.

Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. 2022. ClassActionPrediction: A challenging benchmark for legal judgment prediction of class action cases in the US. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 31–46, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Weijie Shi, Han Zhu, Jiaming Ji, Mengze Li, Jipeng Zhang, Ruiyuan Zhang, Jia Zhu, Jiajie Xu, Sirui Han, and Yike Guo. 2025. LegalReasoner: Step-wised verification-correction for legal judgment reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7297–7313, Vienna, Austria. Association for Computational Linguistics.

Parjanya Aditya Shukla, Shubham Kumar Nigam, Debtanu Datta, Balaramamahanthi Deepak Patnaik, Noel Shallum, Pradeep Reddy Vanga, Saptarshi Ghosh, and Arnab Bhattacharya. 2026. Nyayamind-a framework for transparent legal reasoning and judgment prediction in the indian legal system. *arXiv preprint arXiv:2604.09069*.

Santosh T.Y.S.S and Irtiza Chowdhury. 2025. Fairness beyond performance: Revealing reliability disparities across groups in legal NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24376–24390, Vienna, Austria. Association for Computational Linguistics.

Santosh T.Y.S.S, Mohamed Hesham Elganayni, Stanisław Sójka, and Matthias Grabmair. 2024. Incorporating precedents for legal judgement prediction on European court of human rights cases. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3743–3750, Miami, Florida, USA. Association for Computational Linguistics.

United States Census Bureau. 2016. Frequently occurring surnames from the 2010 census. `https://www.census.gov/data/developers/data-sets/surnames.html`. Accessed 2026-05-26.

Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10882–10893. PMLR.

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*.

Anne L Washington. 2018. How to argue with an algorithm: Lessons from the compas-propublica debate. *The Colorado Technology Law Journal*, 17(1):131.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and 1 others. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Huiyuan Xie, Felix Steffek, Joana De Faria, Christine Carter, and Jonathan Rutherford. 2024. The CLC-UKET dataset: Benchmarking case outcome prediction for the UK employment tribunal. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 81–96, Miami, FL, USA. Association for Computational Linguistics.

Zongyue Xue, Siyuan Zheng, Shaochun Wang, Yiran Hu, Yuxin Yao, Shengran Wang, Haitao Li, Qingyao Ai, Yiqun Liu, Yun Liu, and Weixing Shen. 2025. Justeva: A toolkit to evaluate llm fairness in legal knowledge inference. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pages 6738–6742, New York, NY, USA. Association for Computing Machinery.

Kepu Zhang, Haoyue Yang, Xu Tang, Weijie Yu, and Jun Xu. 2025. Beyond guilt: Legal judgment prediction with trichotomous reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1815–1826, Suzhou, China. Association for Computational Linguistics.

Yuan Zhang, Wanhong Huang, Yi Feng, Chuanyi Li, Zhiwei Fei, Jidong Ge, Bin Luo, and Vincent Ng. 2024. LJPCheck: Functional tests for legal judgment prediction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5878–5894, Bangkok, Thailand. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

A Evaluated Models

Table 4 reports the models used in the evaluation framework, the corresponding Hugging Face checkpoints, model categories, and parameter scales. The evaluated models include general-purpose encoders, legal-domain specialized encoders, and open base and instruction-tuned generative LLMs ranging from 110M to 8B parameters.

Table 4: Language models used in our experiments.

Model	Huggingface Source	Type	Size
BERT	bert-base-uncased	Encoder	110M
LegalBERT	nlpaueb/legal-bert-base-uncased	Legal Encoder	110M
CaseLawBERT	zlucia/custom-legalbert	Legal Encoder	110M
LegalRoBERTa	saibo/legal-roberta-base	Legal Encoder	125M
Llama-3.1-8B-Instruct	meta-llama/Llama-3.1-8B-Instruct	Open LLM instruct	8B
Qwen-2.5-7B-Instruct	Qwen/Qwen2.5-7B-Instruct	Open LLM instruct	7B
Llama-3.1-8B	meta-llama/Llama-3.1-8B	Open LLM base	8B
Qwen-2.5-7B	Qwen/Qwen2.5-7B	Open LLM base	7B
Saul-7B-Instruct-v1	Equall/Saul-7B-Instruct-v1	Legal LLM Instruct	7B

B Experimental Settings

Figure 5 shows the prompt used to instruct the model to predict a binary verdict for a class-action case, with few-shot demonstrations included only in the few-shot setting. In this figure, Grey-boxed text is passed verbatim to the model; italicised slots are filled at inference time. The few-shot block is omitted in the zero-shot and Fine-tuned settings.

```
You are a legal expert. Read the
class-action case below and predict the
verdict. Respond with exactly one digit:
'1' if the plaintiff wins, or '0' if the
plaintiff loses.

[Few-shot only] k labelled demonstrations:
### Case: ⟨case text⟩
### Verdict: ⟨0 or 1⟩

### Case: ⟨test case⟩
### Verdict:
```

Figure 5: Prompt template used across zero-shot, few-shot, and fine-tuned evaluation settings for legal judgment prediction.

Data splits and shared training settings. For all supervised runs, we follow the ClassActionPrediction setup: 70% training, 15% validation, 15% test. Encoder models use learning rate 1×10^{-5}, batch size 16, dropout 0, weight decay 0, softmax output layer, and validation-loss early stopping with patience 2 (max 10 epochs).

Encoder fine-tuning. The BERT-family models use maximum sequence length 512 tokens. The 768-dimensional pooled representation is passed to a task-specific linear classifier for binary win/loss prediction.

Open LLM prompting and fine-tuning. Case text is truncated to a 2,500 token budget. Few-

shot prompting uses $k = 2$ demonstrations with a fixed seed offset by 1000. LoRA fine-tuning uses rank 16, $\alpha = 32$, $dropout = 0.05$, adapters on `q_proj`, `k_proj`, `v_proj`, `o_proj`; learning rate $lr = 5 \times 10^{-5}$, cosine scheduling, warmup ratio of 0.03, 3 epochs, micro-batch of 1, gradient accumulation 16 steps, max sequence length of 3072, FP16, `paged_adamw_8bit` .

Verdict probabilities. For encoder models, $p(win)$ is the softmax probability of the *win* class from the binary classification head. For generative models, it is the probability assigned to the token "1" at the verdict position, renormalized over {"0", "1"}.

Dataset Variants Created for Bias Propagation Experiments For the bias propagation study, four biased training-set variants were derived from the ClassActionPrediction training split: two for gender ('Female Bias' and 'Male Bias') and two for ethnicity ('Black Bias', 'White Bias'). Each variant was constructed deterministically: every training case was assigned the demographic variant text corresponding to its gold verdict (win-tied variant for wins, lose-tied variant for losses), creating a 100% coupling between the demographic cue and the label. Prior to this swap, the training split was subsampled to a 50/50 win/lose balance, ensuring the demographic signal was the sole systematic difference between classes; cases with no valid variant (e.g. corporate plaintiffs) were dropped.

C Evaluation metrics

Counterfactual Flip Rate is formally defined as follows:

$$\mathrm{CFR} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[\hat{y}^{(i)}_{\text{anonym}} \neq \hat{y}^{(i)}_{\text{variant}}\right]$$

where $\hat{y}_{\text{anonym}}$ and $\hat{y}_{\text{variant}}$ denote the predictions for the anonymized case and its demographic variant, respectively. The indicator function $\mathbb{1}[\cdot]$ takes the value 1 when the prediction changes between the two versions of the case and 0 otherwise.

Demographic Parity Gap is defined as:

$$\text{DP-Gap} = \max_{g \in G} P(\hat{y} = 1 \mid g) - \min_{g \in G} P(\hat{y} = 1 \mid g)$$

where G denotes the set of demographic groups.

Propagation score $\tau \in [-0.5, 0.5]$ for a test case c is defined as follows:

$$\tau_c = \frac{\big(p(c_g) - p(c_{anonym})\big) - \big(p(c_{g'}) - p(c_{anonym})\big)}{2}$$

Table 5: Encoder accuracy: paired t-test against base BERT.

Comparison vs. base BERT	Δ Acc.	t	p
BERT-ft	$+0.113$	3.90	***
LegalBERT-ft	$+0.090$	2.80	**
CaseLawBERT-ft	$+0.092$	2.77	**
LegalRoBERTa-ft	$+0.118$	4.15	***
LegalBERT (no-ft)	$+0.005$	0.20	ns
CaseLawBERT (no-ft)	-0.009	-0.20	ns
LegalRoBERTa (no-ft)	-0.012	-1.39	ns

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$; ns = not significant.

where $p(\cdot)$ denotes the predicted probability of a winning verdict. Here, g and g' denote contrasting demographic variants within the same demographic group, such as male and female within the gender category. For each test case c, we construct three variants consisting of an anonymized version c_{anonym} in which all demographic references are removed, a demographic specific version c_g in which references are replaced with the favored demographic group, and a contrasting version $c_{g'}$ in which the same references are replaced with the opposing demographic group. Expanding the two differences, the anonymized term cancels out, so τ_c is reduced to half the difference in predicted winning probability between the favored and contrasting variants. The anonymized version defines the two component shifts, which indicate whether propagation raises the favored group's predicted win probability, lowers the contrasting group's, or both. The propagation score τ is computed as the average of τ_c over all test cases.

D Statistical Significance

Tables 5–8 report the results of paired per-case t-tests for both accuracy and CFR, conducted on the 433 test cases with an individual (non-corporate) representative plaintiff; corporate-plaintiff cases have no personal names or pronouns to perturb.

Encoders (reference: base BERT). All four fine-tuned encoders significantly improve accuracy and increase CFR, except LegalRoBERTa-ft on gender. None of the non-fine-tuned base encoders significantly improves accuracy, while base LegalBERT is already highly sensitive in terms of CFR.

Generative LLMs (reference: Saul-7B-Instruct-zs). Fine-tuned Llama-3.1-8B, Llama-3.1-8B-Instruct, and Saul-7B-Instruct significantly improve accuracy relative to Saul-7B-Instruct-zs. Neither fine-tuned Qwen model produces a significant improvement. Several zero-shot and few-shot

Table 6: Encoder AVG CFR: paired t-test against base BERT. G/ E/ I = Gender/ Ethnicity/ Intersectional.

	Δ CFR			p		
Comparison vs. base BERT	G	E	I	G	E	I
BERT-ft	$+0.058$	$+0.068$	$+0.090$	***	***	***
LegalBERT-ft	$+0.028$	$+0.036$	$+0.047$	***	***	***
CaseLawBERT-ft	$+0.099$	$+0.129$	$+0.128$	***	***	***
LegalRoBERTa-ft	$+0.008$	$+0.013$	$+0.015$	ns	*	*
LegalBERT (no-ft)	$+0.117$	$+0.113$	$+0.155$	***	***	***
CaseLawBERT (no-ft)	$+0.008$	-0.004	-0.001	ns	ns	ns
LegalRoBERTa (no-ft)	-0.003	-0.007	-0.008	ns	*	*

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$; ns = not significant.

Table 7: Generative-LLM accuracy: paired t-test against Saul-7B-Instruct-zs.

Comparison vs. Saul-7B-Instruct-zs	Δ Acc.	t	p
Llama-3.1-8B-zs †	$+0.000$	–	ns
Llama-3.1-8B-fs †	$+0.000$	–	ns
Llama-3.1-8B-ft	$+0.136$	4.16	***
Llama-3.1-8B-Instruct-zs †	$+0.000$	–	ns
Llama-3.1-8B-Instruct-fs	$+0.007$	1.74	ns
Llama-3.1-8B-Instruct-ft	$+0.120$	3.66	***
Qwen-2.5-7B-zs	$+0.018$	0.40	ns
Qwen-2.5-7B-fs	-0.030	-0.73	ns
Qwen-2.5-7B-ft	$+0.018$	0.40	ns
Qwen-2.5-7B-Instruct-zs	$+0.009$	1.16	ns
Qwen-2.5-7B-Instruct-fs	$+0.048$	3.16	**
Qwen-2.5-7B-Instruct-ft	$+0.009$	1.16	ns
Saul-7B-Instruct-fs	$+0.023$	1.39	ns
Saul-7B-Instruct-ft	$+0.132$	4.09	***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$; ns = not significant.
†The model produces predictions identical to Saul-7B-Instruct-zs for every test instance. The paired differences therefore have zero variance, making the paired t-statistic undefined.

models collapse to a constant predictor that produces predictions identical to the reference model; these cases are marked with †. Because the paired differences are zero for every test instance, their paired t-statistics are undefined. For CFR, Saul-7B-Instruct-ft is the least counterfactually sensitive among the fine-tuned generative models (excluding Qwen, whose fine-tuning had no effect) and is the only generative model that approaches the robustness observed for the encoder models.

E Data Preprocessing

From the testing subset of the adopted dataset, we remove a few cases by detecting legally sensitive demographic-specific keywords and expressions that may directly reveal or depend on protected demographic attributes, as well as those where we would be unable to flip the gender descriptors. A pregnancy-leave related complaint, for instance, cannot simply be flipped to a male plaintiff. Further examples include terms related to paternity, alimony, racism, and sexual harassment.

Table 8: Generative-LLM AVG CFR: paired t-test against Saul-7B-Instruct-zs. G/ E/ I = Gender/ Ethnicity/ Intersectional.

	Δ CFR			p		
Comparison vs. Saul-7B-Instruct-zs	G	E	I	G	E	I
Llama-3.1-8B-zs †	$+0.000$	$+0.000$	$+0.000$	–	–	–
Llama-3.1-8B-fs †	$+0.000$	$+0.000$	$+0.000$	–	–	–
Llama-3.1-8B-ft	$+0.047$	$+0.043$	$+0.045$	***	***	***
Llama-3.1-8B-Instruct-zs †	$+0.000$	$+0.000$	$+0.000$	–	–	–
Llama-3.1-8B-Instruct-fs	$+0.002$	$+0.006$	$+0.006$	ns	*	*
Llama-3.1-8B-Instruct-ft	$+0.060$	$+0.085$	$+0.091$	***	***	***
Qwen-2.5-7B-zs	$+0.114$	$+0.110$	$+0.120$	***	***	***
Qwen-2.5-7B-fs	$+0.090$	$+0.097$	$+0.104$	***	***	***
Qwen-2.5-7B-ft	$+0.116$	$+0.109$	$+0.120$	***	***	***
Qwen-2.5-7B-Instruct-zs	$+0.010$	$+0.012$	$+0.014$	*	**	**
Qwen-2.5-7B-Instruct-fs	$+0.022$	$+0.034$	$+0.038$	***	***	***
Qwen-2.5-7B-Instruct-ft	$+0.010$	$+0.011$	$+0.015$	*	**	**
Saul-7B-Instruct-fs	$+0.065$	$+0.062$	$+0.061$	***	***	***
Saul-7B-Instruct-ft	$+0.009$	$+0.011$	$+0.036$	*	**	***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$; ns = not significant.
†The model produces predictions identical to Saul-7B-Instruct-zs for every test instance. The paired differences therefore have zero variance, making the paired t-statistic undefined.

Listing 1: Keyword Patterns Used for Filtering

```
GENDER_PATTERNS = {
    "pregnancy/maternity":
        r"\b(?:pregnan\w*|maternity|prenatal|
    childbirth|nursing mother|"
        r"breastfeed\w*|lactation|miscarriage)\b
    ",

    "paternity/family-leave":
        r"\b(?:paternity|FMLA|family medical
    leave|parental leave)\b",

    "sex/gender discrimination":
        r"\b(?:sex discrimination|gender
    discrimination|sexual harassment|"
        r"gender identity|transgender|misgender\
    w*)\b",

    "sexual violence":
        r"\b(?:rape|sexual assault|sexual abuse|
    sexual battery)\b",

    "reproductive":
        r"\b(?:abortion|contracepti\w*|
    reproductive (?:health|right)\w*|"
        r"IVF|in vitro)\b",

    "title VII/IX (sex)":
        r"\b(?:Title VII|Title IX)\b",

    "marital/spousal status":
        r"\b(?:widow|widower|spousal|alimony|
    divorce\w*)\b",
}

ETHNIC_PATTERNS = {
    "race discrimination":
        r"\b(?:race discrimination|racial
    discrimination|racial slur|"
        r"racism|racist|racially)\b",

    "national origin":
        r"\b(?:national origin|ethnic\w*|
    xenophob\w*)\b",

    "specific race terms":
```

```
        r"\b(?:African[- ]American|Hispanic|
    Latino|Latina|Caucasian|"
        r"Asian[- ]American)\b",

    "civil rights statutes":
        r"\b(?:Section 1981|42 U\.?S\.?C\.?
    1981|Voting Rights Act|"
        r"Fair Housing)\b",

    "religion (often proxied)":
        r"\b(?:religious discrimination|religion
     discrimination)\b",
}
```

F Counterfactual Perturbation of Class Action Cases

In this dataset, Class Actions are represented by an individual labeled "plaintiff" acting on behalf of themself and the class. For instance, a case would state "Cintas employed Plaintiff, Mark Cassingham, as a Technician...". Following that, the claimant party would then be referred to using phrases such as "Plaintiff and the Rule 23 California Class.." or "Plaintiff... on behalf of...".

As such, our counterfactual perturbation process is applied only to the individual or representative plaintiff. This leaves every reference to the class as a collective body untouched. The analysis therefore asks whether changing the perceived demographic identity of the representative plaintiff, while preserving the legal allegations, case facts, and all other information, changes the model's predicted outcome.

G Future Work

We suggest two mitigation strategies. First, counterfactual data augmentation could train models on both the original cases and demographically perturbed versions of the same cases, directly targeting the counterfactual sensitivity identified in our analysis. Second, inference-time debiasing could restrict retrieval to demographically anonymized precedents, thereby separating the effects of demographic information in the query and retrieved cases from the predicted outcome. Evaluating these mitigation strategies is an important next step motivated by the findings of this study.