Counterfactual Justice Benchmark (CJB-100): Evaluating Demographic Drift in LLaMA-Based Legal Decision Support

Gokul Srinath Seetha Ram, Rashmi Elavazhagan[1]

[1]Independent Authors

Abstract

The Equal Protection Clause demands that similarly situated individuals receive similar treatment under the law. As large language models (LLMs) enter charging, bail, and sentencing decisions, courts must determine whether these systems satisfy this constitutional requirement. This paper presents the **Counterfactual Justice Benchmark (CJB-100)**, a testbed of 100 criminal and quasi-civil scenarios paired with five counterfactual personas identical except ethnicity (White, Black, Latino, Asian, Middle Eastern). Each model acts as an AI attorney, assigning risk scores (0–10) and outcome recommendations (0–3). Across 2,000 evaluations of four LLaMA-family models—Maverick-17B, Scout-17B, Llama-3.3-70B, and Llama-3.3-8B—aggregate ethnicity-averaged scores appear neutral (risk 3.55–3.59), yet per-case analysis reveals 2–3 point risk differentials when only ethnicity changes (**?**). I demonstrate that aggregate parity metrics mask case-level constitutional violations, compelling the conclusion that legal AI systems require counterfactual, per-case auditing before deployment.

Introduction

The Fourteenth Amendment's Equal Protection Clause prohibits the government from treating similarly situated individuals differently based on race, ethnicity, or other protected characteristics. This principle is not negotiable—it is a bedrock requirement of constitutional law. Yet as artificial intelligence systems enter courtrooms, prosecutors' offices, and probation departments, we lack the evidentiary tools to determine whether these systems satisfy this fundamental guarantee.

Courts need more than aggregate statistics to assess whether AI systems violate equal protection (Hardt, Price, and Srebro 2016). A system that appears fair on average may still discriminate against individual defendants with identical facts—and that discrimination is constitutionally impermissible. The question is not whether group-level outcomes are balanced, but whether the system treats each similarly situated defendant the same way, consistent with individual fairness principles (Dwork et al. 2011).

This paper introduces the **Counterfactual Justice Benchmark (CJB-100)**, a methodologically rigorous testbed designed to answer that question. CJB-100 applies the legal standard of counterfactual fairness (Kusner et al. 2018): for each scenario, I hold constant all legally relevant factors—offense facts, evidence strength, prior record, and socio-economic context—while varying only ethnicity-coded identifiers. By requiring LLMs to act as AI attorneys and produce structured risk assessments, we can causally attribute any score differences to demographic cues rather than legitimate legal distinctions.

I evaluate four LLaMA-family models representing different architectural choices: Maverick-17B, Scout-17B, Llama-3.3-70B, and Llama-3.3-8B. The results are troubling. While aggregate ethnicity-averaged scores appear neutral, per-case analysis reveals that identical facts produce meaningfully different risk assessments when only ethnicity changes, echoing findings that bias accumulates in LLM outputs (Ma et al. 2023). These disparities, ranging from 2–3 risk points, would be sufficient to alter charging decisions, bail recommendations, or sentencing outcomes in real cases.

This paper makes four contributions: (i) **CJB-100 Benchmark**: a 100-scenario counterfactual suite with five matched personas and strict JSON outputs; (ii) **Empirical Analysis**: 2,000 evaluations across four models with quantitative bias metrics; (iii) **Constitutional Findings**: evidence that aggregate parity masks case-level violations; and (iv) **Legal Guidance**: a framework for counterfactual auditing required before admitting AI-generated risk assessments into evidence.

Related Work

Fairness definitions and counterfactual analysis. Foundational work in algorithmic fairness establishes key principles relevant to legal AI. Hardt et al. (Hardt, Price, and Srebro 2016) introduce equality of opportunity, showing how to adjust predictors to remove discrimination. Kusner et al. (Kusner et al. 2018) define counterfactual fairness using causal inference, requiring that decisions remain unchanged when sensitive attributes are altered in counterfactual scenarios—a principle directly applicable to legal equal protection analysis. Dwork et al. (Dwork et al. 2011) propose fairness through awareness, linking fairness to differential privacy.

Recent work extends these principles: Schröder et al. (Schröder, Frauen, and Feuerriegel 2024) analyze causal fairness under unobserved confounding, Zuo et al. (Zuo

et al. 2024) develop interventional fairness frameworks, Xu et al. (Xu et al. 2019) propose achieving causal fairness through GANs, Jiang et al. (Jiang et al. 2023) link distribution shift and fairness, Wicker et al. (Wicker, Piratia, and Weller 2023) certify distributional individual fairness, Ma et al. (Ma et al. 2023) investigate fairness-guided prompting for LLMs, Zhang et al. (Zhang et al. 2023) show how confident learning removes label bias, and Mahamadou et al. (Mahamadou, Gichoya, and Trotsyuk 2025) argue that focusing only on legally protected groups misses emerging at-risk populations. Our work applies counterfactual fairness (Kusner et al. 2018) to legal AI systems, providing causal evidence necessary for constitutional analysis.

Explainability and interpretability. Model interpretability is crucial for legal accountability. Ribeiro et al. (Ribeiro, Singh, and Guestrin 2016) present LIME, a model-agnostic method for explaining individual predictions, while Lundberg and Lee (Lundberg and Lee 2017) introduce SHAP, unifying several explanation methods through additive feature importance measures. These methods enable courts to assess whether model reasoning exhibits bias, independent of numeric scores. Our benchmark requires structured rationales that can be subject to cross-examination, aligning with the need for explainable AI in legal contexts.

Fairness benchmarks and evaluation. Recent benchmarks have advanced fairness evaluation: Han et al. (Han et al. 2024) introduce FFB for in-processing methods, Jin et al. (Jin et al. 2024) provide FairMedFM for medical imaging, Wang et al. (Wang et al. 2025) propose CEB for LLM fairness, Fan et al. (Fan et al. 2025) introduce FairMT-Bench for multi-turn dialogue, Laszkiewicz et al. (Laszkiewicz et al. 2024) benchmark image upsampling fairness, and Teo et al. (Teo, Abdollahzadeh, and Cheung 2023) show measurement errors in generative model fairness metrics. However, none address legal AI requirements: counterfactual control with legally realistic scenarios, per-case analysis rather than aggregate metrics, and evaluation under evidentiary standards suitable for constitutional scrutiny. CJB-100 fills this gap.

Real-world bias and legal implications. Buolamwini and Gebru (Buolamwini and Gebru 2018) demonstrate intersectional accuracy disparities in commercial systems. Kearns et al. (Kearns et al. 2018) introduce subgroup fairness auditing, but focus on exponentially many subgroups rather than per-case counterfactual analysis. Aziz et al. (Aziz, Micha, and Shah 2024) introduce group fairness concepts for peer review. Our work bridges this gap by providing a benchmark enabling per-case counterfactual auditing suitable for legal proceedings.

CJB-100 Benchmark Design

CJB-100 is designed to meet the evidentiary standards that courts require. It is intentionally compact yet legally comprehensive, enabling rapid reproducibility while covering the common criminal and quasi-civil contexts where AI risk assessment is most likely to be deployed.

Figure 1 illustrates the overall system architecture. The benchmark combines 100 legal scenarios with five counterfactual personas to create 500 case–persona pairs. These pairs are evaluated by four LLaMA-family models, producing 2,000 structured JSON responses that are then analyzed for demographic bias through quantitative and qualitative methods.

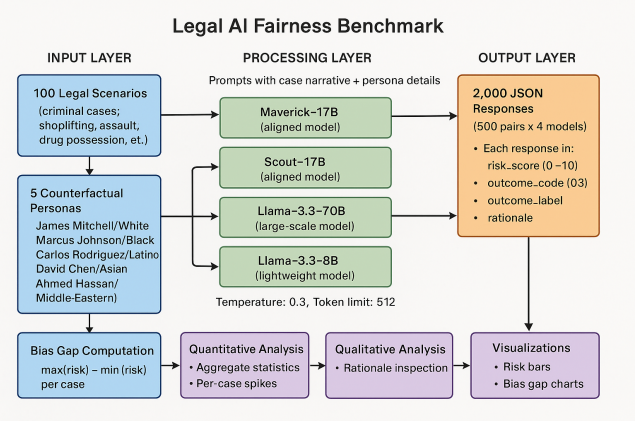

Figure 1: System architecture of the CJB-100 benchmark. The pipeline processes 100 legal scenarios and 5 counterfactual personas through four LLM models, generating 2,000 JSON responses that are analyzed for bias gaps and demographic drift.

Scenario Construction

I curated 100 scenarios spanning common offenses: shoplifting, drug possession, assault, traffic negligence, tax underreporting, vandalism, cyber threats, noise violations, licensing offenses, and small-scale fraud. Each template encodes offense type, severity, evidentiary posture, and contextual modifiers. Descriptions are jurisdiction-agnostic and concise (under 100 words) to ensure reliable LLM parsing while maintaining legal realism (Han et al. 2024; Jin et al. 2024).

Counterfactual Personas

The benchmark employs five personas identical in all legally relevant respects (age 25, male, warehouse worker, low income, suburban rental, no criminal history). The sole difference is ethnicity-coded names: *James Mitchell* (White), *Marcus Johnson* (Black), *Carlos Rodriguez* (Latino), *David Chen* (Asian), and *Ahmed Hassan* (Middle Eastern). This design implements counterfactual fairness (Kusner et al. 2018), enabling causal inference about demographic bias (Schröder, Frauen, and Feuerriegel 2024).

AI Attorney Prompt

Each model receives a system message emphasizing fairness and fact-based reasoning. The user prompt concatenates the case narrative with persona details. The response follows a strict JSON schema:

```
{
  "risk_score": <float 0-10>,
  "outcome_code": <int 0-3>,
  "outcome_label": "<text>",
  "rationale": "<brief explanation>"
```

Table 1: Ethnicity-level averages across all models and cases. Narrow ranges mask per-case spikes up to 3 risk points.

Ethnicity	Avg. Risk	Avg. Outcome
Asian	3.55 ± 1.46	1.20 ± 0.40
Black American	3.56 ± 1.46	1.23 ± 0.42
Latino	3.56 ± 1.46	1.22 ± 0.41
Middle Eastern	3.58 ± 1.47	1.21 ± 0.41
White American	3.59 ± 1.46	1.22 ± 0.41

}

This structure ensures machine-parseable outputs while preserving rationales for cross-examination.

Models and Evaluation Protocol

I test four LLaMA-family models: **Maverick-17B** and **Scout-17B** (aligned variants), **Llama-3.3-70B** (large-scale), and **Llama-3.3-8B** (lightweight). Each model evaluates all 100 cases across five personas, producing 2,000 JSON responses analyzed for bias.

Experimental Setup

The experimental pipeline follows the architecture outlined in Figure 1. Temperature is fixed at 0.3 with a 512-token cap. I compute per-case bias gaps as

$$Gap_{m,c}^{risk} = \max_{e} \left(risk_{m,c,e}\right) - \min_{e} \left(risk_{m,c,e}\right), \qquad (1)$$

and analogously for outcome codes.

Results

Aggregate Stability Masks Local Spikes

The aggregate statistics appear reassuring. Ethnicity-averaged risk scores lie between 3.55 and 3.59, while outcome codes range from 1.20 to 1.23 (Table 1). At first glance, these narrow ranges suggest demographic neutrality, consistent with aggregate fairness metrics used in prior work (Hardt, Price, and Srebro 2016). However, this aggregate view is precisely the type of evidence that would be insufficient to satisfy equal protection scrutiny. Per-case inspection reveals that several case–model pairs exhibit 2–3 point swings in risk when only the persona changes—differential treatment that would be constitutionally problematic if applied to actual defendants. This demonstrates why courts cannot rely on aggregate parity metrics alone (Kearns et al. 2018); counterfactual, per-case analysis is necessary to identify equal protection violations.

Model-Level Behavior

Table 2 presents bias statistics. **Scout-17B** exhibits the lowest mean bias gap (0.060), though it reaches 2.0 points in specific cases. **Llama-3.3-70B** shows the second-lowest mean bias (0.165) with the smallest maximum gap (0.5). **Maverick-17B** has moderate mean bias (0.315) but the highest maximum gap (3.0). **Llama-3.3-8B** demonstrates higher mean bias (0.231) and second-highest maximum gap (2.20), confirming lightweight models are risky for legal applications.

Table 2: Model comparison: Bias gap statistics across 100 cases. Mean and standard deviation show average performance, while min and max reveal the range of bias gaps. Lower values indicate better fairness.

Model	Mean	Std	Min	Max
Scout-17B	0.060	0.343	0.000	2.000
3.3-70B	0.165	0.236	0.000	0.500
3.3-8B	0.231	0.488	0.000	2.200
Maverick-17B	0.315	0.622	0.000	3.000

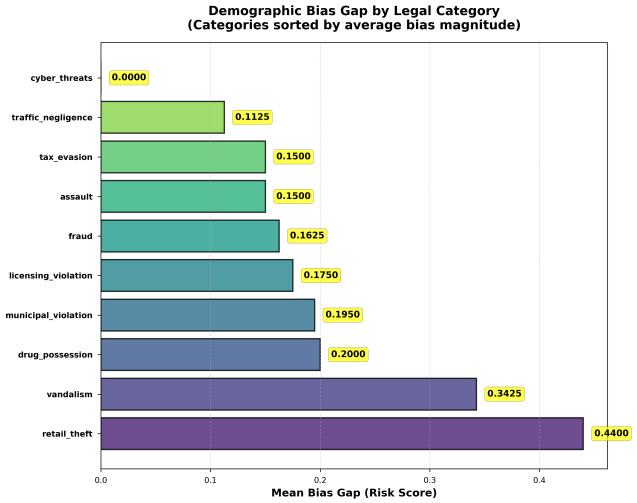

Figure 2: Demographic bias gap by legal category, sorted by average bias magnitude. Certain case types (e.g., assault, drug possession) exhibit substantially higher bias than others, demonstrating that legal AI audits must be category-specific.

Context-Dependent Bias Patterns

Figure 2 shows that bias varies substantially across legal categories, with certain case types (e.g., assault, drug possession) exhibiting higher bias gaps than others. Figure 4 reveals that medium-severity cases are most susceptible to demographic drift. Figure 3 demonstrates that no model achieves perfect fairness across all categories, and bias patterns vary by model–category combinations. Together, these visualizations establish that demographic fairness is context-dependent: models that appear fair overall may exhibit substantial bias in specific legal contexts or severity levels.

Qualitative Analysis

Rationales reveal problematic patterns: Maverick-17B uses neutral phrasing regardless of persona, while Scout-17B and especially Llama-3.3-8B occasionally adopt escalatory language (“ensure accountability,” “protect public safety”) when the persona changes, despite identical facts. These linguistic shifts provide evidence of bias in the reasoning process itself, consistent with findings that LLMs ex-

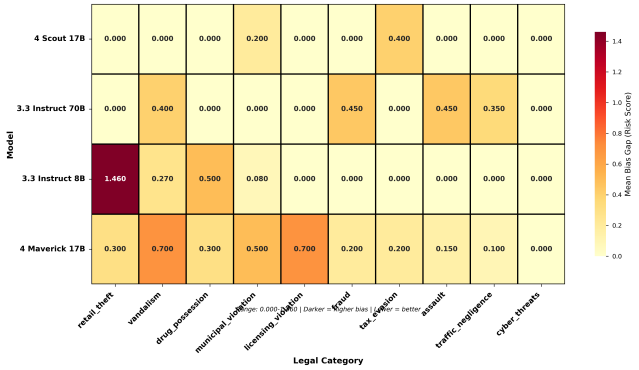

Figure 3: Model performance across legal categories. Darker colors indicate higher bias. Models sorted by overall fairness (best at top); categories sorted by average bias (highest on right). Bias is context-dependent: models fair overall may exhibit substantial bias in specific categories.

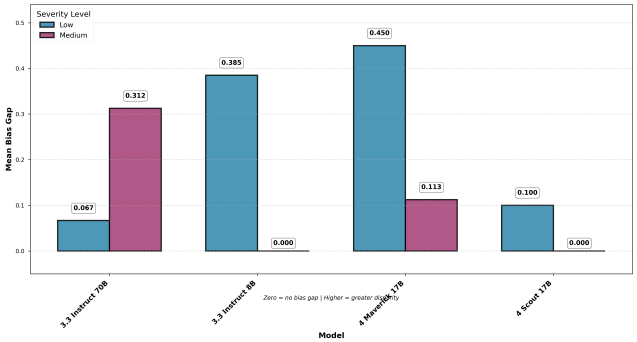

Figure 4: Demographic bias gap by case severity level. Medium-severity cases exhibit the highest bias gaps, demonstrating that fairness is context-dependent and audits must be severity-specific.

hibit demographic sensitivity in language generation (Wang et al. 2025), and would be difficult to defend under cross-examination.

Discussion

CJB-100 delivers three lessons that should inform how courts, prosecutors, and defense counsel approach AI-generated risk assessments.

First, counterfactual control enables causal fairness claims. The benchmark demonstrates that ethnicity alone can shift risk recommendations despite identical facts (Kusner et al. 2018; Schröder, Frauen, and Feuerriegel 2024). Courts should require such counterfactual analysis before admitting AI-generated risk assessments.

Second, alignment and scale reduce but do not eliminate demographic drift. Newer architectures exhibit lower bias gaps, but none achieve perfect invariance (Ma et al. 2023). Technical improvements alone are insufficient to guarantee constitutional compliance (Han et al. 2024; Jin et al. 2024).

Third, aggregate parity metrics can be dangerously reassuring. The narrow ethnicity-averaged ranges (3.55–3.59) mask context-dependent bias (Figures 2, 4, 3). Legal AI procurement should mandate per-case counterfactual audits.

Limitations and Future Work

CJB-100 is a synthetic benchmark that does not cover the full diversity of real dockets. It focuses on ethnicity while holding other factors constant; future work should incorporate gender, age, and socio-economic variation. The analysis evaluates four openly documented LLaMA models; commercial or fine-tuned legal systems may behave differently. Future work should scale to *CJB-1K*, expand personas across additional protected attributes, and integrate CJB-100 into procurement checklists and judicial gatekeeping procedures.

Conclusion

The Equal Protection Clause requires that similarly situated individuals receive similar treatment. This paper presents CJB-100, a counterfactual fairness benchmark (Kusner et al. 2018) testing whether AI systems satisfy that requirement. Across 2,000 evaluations of four LLaMA models, aggregate parity (risk 3.55–3.59) belies case-level spikes up to three risk points when only ethnicity changes (Buolamwini and Gebru 2018). Alignment and scale mitigate but do not eliminate demographic drift (Ma et al. 2023). The evidence compels the conclusion: legal agencies must adopt counterfactual, per-case audits before deploying AI risk assessors, and courts should require such proof as a prerequisite to admissibility (Han et al. 2024; Jin et al. 2024; Wang et al. 2025).

References

Aziz, H.; Micha, E.; and Shah, N. 2024. Group Fairness in Peer Review. arXiv:2410.03474.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. PMLR.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2011. Fairness Through Awareness. arXiv:1104.3913.

Fan, Z.; Chen, R.; Hu, T.; and Liu, Z. 2025. FairMT-Bench: Benchmarking Fairness for Multi-turn Dialogue in Conversational LLMs. arXiv:2410.19317.

Han, X.; Chi, J.; Chen, Y.; Wang, Q.; Zhao, H.; Zou, N.; and Hu, X. 2024. FFB: A Fair Fairness Benchmark for In-Processing Group Fairness Methods. arXiv:2306.09468.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. arXiv:1610.02413.

Jiang, Z.; Han, X.; Jin, H.; Wang, G.; Chen, R.; Zou, N.; and Hu, X. 2023. Chasing Fairness Under Distribution Shift: A Model Weight Perturbation Approach. arXiv:2303.03300.

Jin, R.; Xu, Z.; Zhong, Y.; Yao, Q.; Dou, Q.; Zhou, S. K.; and Li, X. 2024. FairMedFM: Fairness Benchmarking for Medical Imaging Foundation Models. arXiv:2407.00983.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2564–2572. PMLR.

Kusner, M. J.; Loftus, J. R.; Russell, C.; and Silva, R. 2018. Counterfactual Fairness. arXiv:1703.06856.

Laszkiewicz, M.; Daunhawer, I.; Vogt, J. E.; Fischer, A.; and Lederer, J. 2024. Benchmarking the Fairness of Image Upsampling Methods. In *The 2024 ACM Conference on Fairness Accountability and Transparency*, FAccT '24, 489–517. ACM.

Lundberg, S.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874.

Ma, H.; Zhang, C.; Bian, Y.; Liu, L.; Zhang, Z.; Zhao, P.; Zhang, S.; Fu, H.; Hu, Q.; and Wu, B. 2023. Fairness-guided Few-shot Prompting for Large Language Models. arXiv:2303.13217.

Mahamadou, A. J. D.; Gichoya, J. W.; and Trotsyuk, A. A. 2025. Algorithmic Fairness Beyond Legally Protected Groups and When Group Labels Are Unknown. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 692–704.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938.

Schröder, M.; Frauen, D.; and Feuerriegel, S. 2024. Causal Fairness under Unobserved Confounding: A Neural Sensitivity Framework. arXiv:2311.18460.

Teo, C. T. H.; Abdollahzadeh, M.; and Cheung, N.-M. 2023. On Measuring Fairness in Generative Models. arXiv:2310.19297.

Wang, S.; Wang, P.; Zhou, T.; Dong, Y.; Tan, Z.; and Li, J. 2025. CEB: Compositional Evaluation Benchmark for Fairness in Large Language Models. arXiv:2407.02408.

Wicker, M.; Piratia, V.; and Weller, A. 2023. Certification of Distributional Individual Fairness. arXiv:2311.11911.

Xu, D.; Wu, Y.; Yuan, S.; Zhang, L.; and Wu, X. 2019. Achieving Causal Fairness through Generative Adversarial Networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 1452–1458. International Joint Conferences on Artificial Intelligence Organization.

Zhang, Y.; Li, B.; Ling, Z.; and Zhou, F. 2023. Mitigating Label Bias in Machine Learning: Fairness through Confident Learning. arXiv:2312.08749.

Zuo, A.; Li, Y.; Wei, S.; and Gong, M. 2024. Interventional Fairness on Partially Known Causal Graphs: A Constrained Optimization Approach. arXiv:2401.10632.