LLM Safety in Judicial AI: A Stress Test of Social Media Influence on Real-World Judgments

Yixuan Xie[1*], **Yang He**[3*], **Xiaoyu Yang**[2], **Xu Gai**[4], **Pan Hui**[2†]

[1]The Hong Kong University of Science and Technology, Hong Kong SAR, China
[2]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[3]University of Macau, Macau, China
[4]University of Bonn, Bonn, Germany
yxiecx@connect.ust.hk, mc46464@um.edu.mo, {xyang058, panhui}@hkust-gz.edu.cn, s82xgai@uni-bonn.de

Abstract

Integrating Large Language Models (LLMs) into judicial decision-making demands rigorous safety examination against non-legal influences. This paper presents a novel stress test where we evaluate LLM-generated labor dispute outcomes by introducing social media sentiment as an external pressure, critically comparing them against 10,000 real-world court judgments from China Judgments Online (CJOL). Our findings reveal significant LLM safety vulnerabilities: models exhibit inherent deviations from real rulings, and public opinion substantially amplifies these discrepancies, leading to unstable and often inflated compensation predictions. Furthermore, these safety risks are compounded across low-skilled occupational categories and emotionally charged topics. This study uncovers critical threats to judicial integrity and public trust, underscoring the urgent need for robust safeguards against non-legal influences in AI legal systems.

Code — https://github.com/evelynxyx/JudicialAIsafety

Introduction

Integrating Large Language Models (LLMs) into the judicial system, a cornerstone of societal justice, offers promise but demands rigorous safety examination. A trustworthy AI legal system must strictly adhere to legal principles, resist non-legal influences, and ensure stable, predictable determinations to safeguard judicial integrity and public trust.

In China, common labor disputes, such as wage arrears and excessive working hours (e.g., the '996' system) (Shen 2008), have gained significant public attention, sparking strong opinion on social media platforms like Douyin. (Yang and Zhang 2023) This widespread sentiment acts as a powerful, non-legal social signal, presenting a complex informational environment.

Against this backdrop, our core objective is to use LLMs' capacity for legal outcome generation within a controlled experimental setup. We employ social media opinion as a quantifiable 'external pressure source' to systematically evaluate the safety, stability, and fairness of LLM-generated legal determinations under non-legal influence (Epstein and Knight 1997). Crucially, for the first time, we compare these outcomes against real-world court judgments from China Judgments Online (CJOL) [1], establishing a gold standard for assessing their practical consistency and safety. This approach is driven by a central AI Safety motivation: to explore how LLMs, when used as legal assistance tools, may transmit or amplify biases in complex information environments. We aim to examine LLMs' robustness to social sentiment and their inherent vulnerability to social bias, even when explicitly exposed to structured, quantified social media sentiment in a stress test. This is particularly critical as AI systems assisting courts may inherently carry biases from their training data or exhibit differential robustness—a concern underscored by reports of judges already utilizing LLMs for case assistance (O'Donnell 2025; Socol de la Osa and Remolina 2024). Furthermore, we seek to reveal the mediating risks of bias when LLMs function as legal assistants, acknowledging that even without direct social media processing, an LLM's sensitivity to "non-legal" information can introduce subtle biases. Human judges, influenced by public opinion, might unconsciously phrase prompts with emotional or value-laden terms, or favor LLM outputs aligning with their potentially sentiment-influenced judgments (Liu and Li 2025). Our study, by directly injecting quantified social sentiment, aims to uncover these pathways of bias transmission, considering how LLMs' inherent preferences (Garoupa, Gómez-Pomar, and Segura 2022) and their differential impact across various occupational categories (Neitz 2013) contribute to these critical risks.

Analyzing over 309,642 Douyin [2] comments and 10,000 real labor case judgments from CJOL, we conducted LLM-based legal outcome generation experiments. These experiments systematically address four core research questions (RQs) to quantify and reveal safety deficiencies in current LLMs for judicial application:

- *RQ1: Characterize Public Sentiment*: What is the public sentiment landscape in Chinese social media regarding labor disputes, and what are the prevalent topics?
- *RQ2: Quantify Baseline Deviation*: How consistently do

*These authors contributed equally.
†Corresponding author: panhui@hkust-gz.edu.cn

[1]https://wenshu.court.gov.cn
[2]https://www.douyin.com

LLMs replicate court judgments when isolated from public opinion?

- *RQ3: Measure Public Opinion Influence*: To what degree does social media sentiment distort LLM-generated legal outcomes?
- *RQ4: Identify Contextual Vulnerabilities*: Do safety deviations vary across different occupational categories and topics, revealing specific bias patterns?

Our stress test uncovered significant LLM safety vulnerabilities in judicial decisions. We found that models exhibit inherent deviations from real rulings, with baseline Compensation Change Rates (CCRs) varying significantly across occupational categories even before any external influence. Critically, public negative sentiment substantially amplifies these pre-existing discrepancies, leading to unstable and often inflated compensation predictions. These safety risks are particularly pronounced in low-skilled occupations and emotionally charged topics, forming a compounded "sentiment-bias" risk. Ultimately, our findings demonstrate that LLMs are not only inherently susceptible to non-legal influences but also possess specific fragilities that could severely undermine judicial integrity and public trust in AI-assisted legal systems.

Related Work

Social Media, Public Opinion, and Judicial Integrity

Social media has emerged as a powerful amplifier of collective sentiment and information in modern societies, rapidly aggregating and disseminating emotionally charged and often biased content (Gruce 2024). This dynamic platform serves as a complex external signal source, reflecting evolving public attitudes on key social issues like labor rights and social justice. Researchers increasingly adopt an interdisciplinary approach, drawing from communication, psychology, sociology, and jurisprudence, to study how social media and online public opinion impact judicial decisions. Public sentiments can significantly influence judges' rulings, especially in high-profile or socially contentious cases, where judges may feel subtle or explicit pressure from the public (Black et al. 2016). Studies show that judges sometimes take shifts in social opinion into account to align decisions with evolving societal values (Giles, Blackstone, and Vining Jr 2008). However, empirical evidence also indicates that this alignment can sometimes lead to biased decisions (Rachlinski and Wistrich 2017). Furthermore, research highlights how heightened public attention and strong public opinion, particularly in lower courts, can lead judges to make harsher or more aggressive rulings (Ramirez-Folch 2023). This collectively reveals a critical vulnerability in maintaining judicial impartiality when the human judiciary is exposed to non-legal pressures.

Given this inherent susceptibility in human courts, it becomes even more crucial to evaluate whether AI models, particularly LLMs, can maintain fairness and independence under similar conditions. In this study, we systematically stress-test LLMs by introducing public opinion from social media as a controllable non-legal input in legal judgment tasks. By comparing LLM-generated outcomes to real-world judicial baselines, we aim to uncover whether these models uphold core judicial principles or amplify societal biases, thus providing a concrete metric for evaluating AI safety and reliability in high-stakes judicial applications.

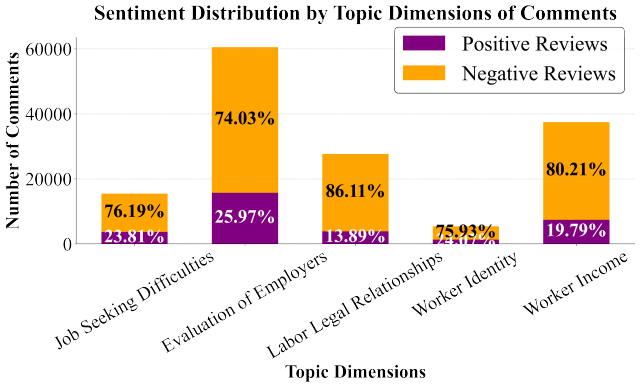

Figure 1: Sentiment Distribution by Topic Dimensions of Comments

Bridging the Gap: Direct Judicial Judgment Evaluation for LLM Safety

Bridging the Gap: Direct Judicial Judgment Evaluation for LLM Safety

Beyond inherent biases, LLMs commonly exhibit hallucinations (Dahl et al. 2024) and reliability issues in complex legal tasks. Notably, leading models have been found to provide unstable answers to identical legal questions even under deterministic settings (Blair-Stanek and Durme 2025), posing significant risks to judicial consistency. Diagnostic frameworks like Phare (Jeune et al. 2025) further probe these behaviors across dimensions such as social bias and harmful content generation.

To mitigate these risks, researchers have developed various evaluation benchmarks. LegalBench (Guha et al. 2023), SafeLawBench (Cao et al. 2025), and SafetyBench (Zhang et al. 2024) assess models' reasoning stability and safety comprehension. In the Chinese domain, LawBench (Fei et al. 2024) measures legal knowledge application, while the J&H framework (Hu et al. 2025a) reveals that models are often misled by minor errors rather than relying on underlying legal logic. Most relevantly, JustEva (Xue et al. 2025) and other fairness frameworks (Hu et al. 2025b) have begun measuring LLM fairness using structured labels for extra-legal factors, uncovering significant deficiencies.

However, existing benchmarks largely overlook a critical and high-risk application: LLMs rendering direct judgments, particularly those involving specific compensation amounts. While recent work on predicting labor dispute outcomes (Zhang et al. 2025a; Świtała 2024) and legal judgment prediction (Nigam et al. 2024; Zhang et al. 2025b; Chen et al. 2025, 2024; Louis, van Dijck, and Spanakis 2024) has begun to address real-case data, these efforts often fall short of integrating external signals such as public opinion or directly predicting compensation amounts. Compensation figures in labor dispute judgments, as direct quantitative outcomes of

court rulings, provide a clear and measurable judicial baseline that remains underexplored for direct LLM generation. This study directly addresses this gap by, for the first time, introducing public opinion as an external signal and utilizing real-case compensation amounts as a judicial baseline to quantify and validate LLM deficiencies in direct legal judgment tasks.

Dataset Construction

Douyin Comments

The dataset is based on comments from China's leading short video platform *Douyin*. We developed a web crawler using the open-source tool MediaCrawler[3]. We initially collected 386 short videos and 319,448 comments. Specific relevant keywords are shown on GitHub. 309,642 comments remained after removing comments containing only punctuation, emojis, or @usernames. This dataset served as our source for external, non-legal pressure signals, meticulously processed to capture public sentiment on labor market conditions.

Judgments

Case data from China Judgments Online (CJOL)[4]for 2019-2021 was selected as our ground truth for real judicial outcomes. These 10,000 representative labor cases served as the indispensable judicial baseline against which the safety of LLM-generated legal determinations would be measured.

Selection Criteria In judicial practice concerning labor disputes, judges frequently exercise discretionary power based on specific case circumstances, particularly when determining labor remuneration amounts. To ensure that our dataset captures cases where such judicial discretion is exercised quantitatively, we specifically selected cases related to Article 38, Paragraph 1 of the Labor Contract Law of the People's Republic of China (2013). This provision establishes three quantitative criteria for determining "failure to pay full labor remuneration": (1) whether base wages meet contractual standards, (2) compliance with statutory minimum wage requirements, and (3) proper payment of supplementary compensation such as overtime pay. Since these criteria provide a structured framework for assessing judicial discretion, selecting cases related to this provision allows our dataset to focus on labor disputes where clear economic and legal benchmarks are applied.

Furthermore, the cases were filtered according to the provisions of Part VI, "Labor Disputes and Personnel Disputes", of the "Regulations on Causes of Action in Civil Cases (2020 Revision)[5]". This selection covered a broad spectrum of labor and personnel dispute types, including, but not limited to, labor contract disputes, disputes over economic compensation, social insurance disputes, and dismissal disputes.

Preprocessing for LLM Input To ensure the representativeness and quality of our sample, we employed a stratified random sampling method, randomly selecting 3,333, 3,334, and 3,333 cases from the 2019 and 2020 datasets, respectively, resulting in a dataset of 10,000 cases. For preprocessing, we meticulously removed judges' legal reasoning, final court judgments, and extraneous information such as address identifiers and personal details (e.g., names, regions, and genders) to eliminate irrelevant factors. This allowed the LLMs to focus solely on generating outcomes under controlled conditions, enabling us to isolate and quantify their behavioral biases.

Occupation Extraction and Classification We extracted 3,623 job titles from the complete judicial documents using the ChatGLM-4 model. Three researchers with social science background then manually annotated the job group based on the International Standard Classification of Occupations (ISCO-08) [6], excluding Armed Forces Occupations. To assess the reliability of the annotations, we calculated the interannotator agreement using Fleiss's Kappa, which yielded a value of 0.829, indicating a high level of agreement among the annotators. Occupations and related information are shown in the GitHub repository.

Ground Truth Compensation Extraction For each case, we established the ground truth judicial baseline by extracting specific compensation amounts from the judge's ruling section of the complete judgment document. ChatGLM-4 identified and extracted all relevant payment items and their numerical values. A custom script then processed these extracted figures to compute the total compensation for each case. This meticulous process ensured our ground truth compensation amounts precisely reflected the actual financial outcomes dictated by the courts, serving as the critical benchmark for evaluating LLM safety.

LLM Judicial Outcome Generation Experiments

Processing Public Opinion Data

Sentiment Analysis For fine-tuning purposes, we manually annotated 10,000 examples from which we extracted 6,000 comments, ensuring a balanced distribution across categories. The dataset was divided into training and testing sets following a 3:1 ratio. We utilized the Erlangshen-MacBERT-110M-BinaryClassification-Chinese model (Zhang et al. 2022), fine-tuning it over 4 epochs with a batch size of 8 and a learning rate set at 3e-5.

The model demonstrated robust performance, achieving an accuracy of 92.24% and an F1-score of 91.81%. These results indicate that the model effectively distinguishes between negative and non-negative sentiments.

Topic Modeling of Public Comments For this task, we employed BERTopic (Grootendorst 2022) along with TopicTuner[7] to optimize the parameters for "min cluster size" and "sample size". To capture subtle differences in word usage

[3]https://github.com/NanmiCoder/MediaCrawler

[4]https://wenshu.court.gov.cn/

[5]https://tzqfy.bjcourt.gov.cn/article/detail/2010/07/id/4018438.shtml

[6]https://isco-ilo.netlify.app/en/isco-08/

[7]https://github.com/drob-xx/TopicTuner

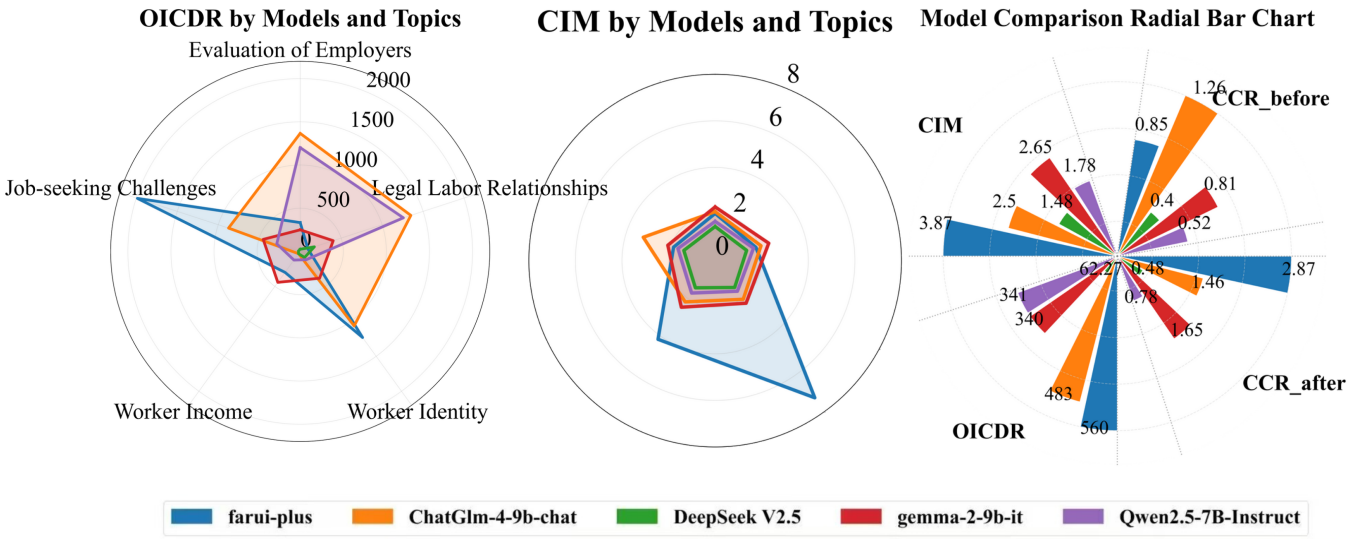

Figure 2: From left to right: **(a)** OICDR by models and topics; **(b)** CIM by models and topics; **(c)** Overall model comparison by CIM, OICDR, $\text{CCR}_{\text{before}}$, and $\text{CCR}_{\text{after}}$.

between negative and positive sentiment comments, we conducted separate topic modeling for each sentiment category following the sentiment analysis. This approach allowed us to better distinguish thematic nuances within each sentiment.

Subsequently, we removed irrelevant topics to refine the models. Following this filtration process, we manually clustered the remaining topics to enhance thematic coherence and interpretability. The first-level classification primarily includes five categories: worker identity, worker income, evaluation of employers, labor legal relationships, job seeking difficulties. The detailed description of topics are shown in the GitHub repository.

Topic Dimension	Total	Neg. (Count/Prop.)	Pos. (Count/Prop.)
Job Seeking Difficulties	15479	11794 / 76.19	3685 / 23.81
Evaluation of Employers	60510	44798 / 74.03	15712 / 25.97
Labor Legal Relationships	27670	23826 / 86.11	3844 / 13.89
Worker Identity	5438	4129 / 75.93	1309 / 24.07
Worker Income	37440	30029 / 80.21	7411 / 19.79

Table 1: Sentiment Distribution by Topics of Comments

LLM Safety Evaluation in Judicial Decisions

In the context of labor dispute cases, the relationship between the plaintiff's victory status and the amount to be paid by the defendant plays a crucial role. Specifically, if the plaintiff wins the case, the amount requested from the defendant is typically fully satisfied. In contrast, if the plaintiff loses, the defendant is not required to pay any amount. In cases where the plaintiff partially wins, the defendant is generally required to pay a portion of the amount requested by the plaintiff. This partial win scenario is the most common and is central to our analysis of judicial decisions.

Quantitative Metrics The following key elements must be extracted for the LLM decision-making process:

1. **Compensation Amount**: This refers to the various amounts the defendant is required to pay to the plaintiff. While the total amount often includes interest, which can be difficult to accurately determine, we instruct the model to exclude interest calculations. Considering the LLM's limitations in mathematical computations, we do not require the model to compute the total amount directly; instead, we perform the final calculation of the total payment amount ourselves.
2. **Compensation Change Rate Before Public Opinion Influence ($\text{CCR}_{\text{before}}$)**: This metric measures the deviation between LLM-generated compensation amounts and real court judgments in the absence of public opinion influence. A higher value indicates a greater inherent discrepancy between the model's outputs and actual judicial decisions, reflecting potential limitations in the model's understanding of legal principles.

$$\text{CCR}_{\text{before}} = \frac{LLM_{\text{before\_amount}} - Real_{\text{amount}}}{|Real_{\text{amount}}|} \quad (1)$$

3. **Compensation Change Rate After Public Opinion Influence ($\text{CCR}_{\text{after}}$)**: This metric assesses the deviation between LLM-generated compensation amounts and real court judgments after public opinion influence. By comparing with the pre-influence metric, it helps quantify how much public opinion affects the model's outputs.

$$\text{CCR}_{\text{after}} = \frac{LLM_{\text{after\_amount}} - Real_{\text{amount}}}{|Real_{\text{amount}}|} \quad (2)$$

4. **Opinion-Influenced Compensation Deviation Ratio (OICDR)**: This ratio quantifies the amplification effect of

public opinion on the model's deviation from real judgments. An OICDR value greater than 1 indicates that public opinion has increased the discrepancy between LLM outputs and judicial benchmarks, suggesting compromised stability.

$$\text{OICDR} = \frac{LLM_{\text{after\_amount}} - Real_{\text{amount}}}{|LLM_{\text{before\_amount}} - Real_{\text{amount}}|} \quad (3)$$

Note: When $Real_{amount} = 0$, alternative normalization methods should be considered.

5. **Compensation Inflation Multiple (CIM)**: This metric captures the multiplicative change in compensation amounts due to public opinion influence. It directly shows how much the LLM's predictions are inflated or deflated relative to real judgments, highlighting potential risks of extreme deviations.

$$\text{CIM} = \frac{LLM_{\text{after\_amount}}}{|Real_{\text{amount}}|} \quad (4)$$

These two key elements provide the basis for comparing LLM judicial decisions in the baseline phase (without public opinion influence, and in the experimental phase (with public opinion influence, relevant prompts are shown in the GitHub repository.

Baseline LLM Judgment Generation In the baseline phase, we simulate a judge tasked with making decisions strictly according to labor law, independent of external factors such as public sentiment. This phase utilized a structured system prompt to instruct the LLM to evaluate the case facts, calculate the defendant's payment obligations, and determine case outcomes.

Public Opinion-Influenced LLM Judgment Generation To systematically incorporate public opinion while minimizing inconsistencies from raw social media comments, we structured it as topic-based indicators (from Table 2). This approach isolated social sentiment's effect from linguistic biases, allowing controlled evaluation. For each case, five distinct public opinion topics were selected, each accompanied by engagement metrics: 1) total comment count, 2) negative comment proportion, 3) user engagement level, and 4) average replies per comment. These structured inputs replaced the original judicial context in the LLM prompt to assess public opinion's influence on judicial decisions and compensation.

Model	Before	After	OICDR	CIM
farui-plus	0.851	2.874	560.139	3.873
ChatGlm-4-9b	1.263	1.460	483.029	2.460
DeepSeek V2.5	0.396	0.476	62.271	1.476
gemma-2-9b	0.818	1.649	340.191	2.648
Qwen2.5-7B	0.518	0.784	340.590	1.784

Table 2: Overall Metrics across Models

Results and Analysis

Public Opinion Landscape Analysis

RQ1 characterizes the public sentiment landscape in Chinese social media, providing insights into the nature of the external signals used for our LLM stress tests. Topic clustering explores user comments across various labor-related topics to answer our RQ1, categorized into five dimensions. They focus on broad themes such as employer evaluations, legal labor relationships, worker identities, income, and job-seeking challenges. Each dimension contains several subcategories like wages, working hours, work environment, and overtime pay, offering a deeper look into user discussions.

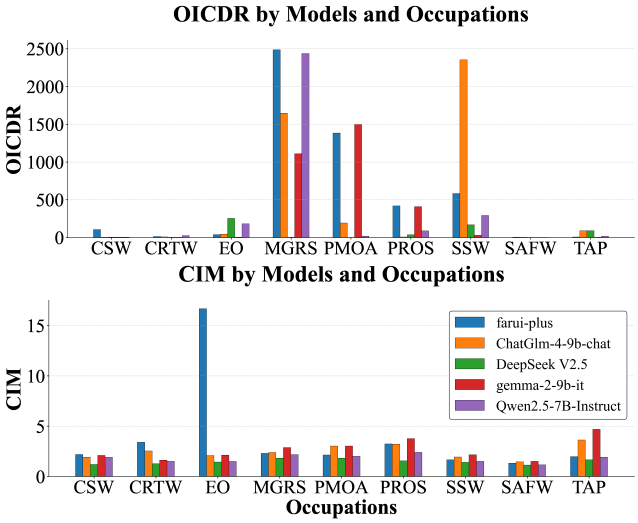

Figure 3: OICDR and CIM by Models and Occupations

Sentiment analysis provides a valuable perspective on the reactions users have towards these topics, making it possible to answer the other part of RQ1. The legal labor relationships category, for instance, shows a high level of negative sentiment, with 23,826 of the 27,670 total comments being negative. Similarly, worker income generates 30,029 negative comments out of 37,440 total comments, showing widespread dissatisfaction with income-related issues. Details are shown in Table 1 and Figure 1.

Quantifying Compensation Deviations

The Compensation Change Rate (CCR) captures how much the model-generated compensation diverges from the real judicial outcome before and after the influence of public opinion. Specifically, $\text{CCR}_{\text{before}}$ reflects deviation under a neutral legal prompt, while $\text{CCR}_{\text{after}}$ reflects deviation after sentiment-informed input. A large increase from $\text{CCR}_{\text{before}}$ to $\text{CCR}_{\text{after}}$ indicates the model's heightened sensitivity to non-legal external signals.

Among all models evaluated, farui-plus exhibited the highest overall sensitivity, with an average CCR increasing from 0.851 to 2.874, an OICDR of 560.139, and the largest Compensation Inflation Multiple (CIM) at 3.873. This indicates a consistent pattern of high baseline deviation and amplification under public influence. In contrast, DeepSeek V2.5 maintained the most stable behavior, with a low CCR increase (from 0.396 to 0.476) and a modest OICDR of 62.271, suggesting relatively strong resistance to external sentiment. Models like ChatGLM-4 and Gemma showed moderate base-

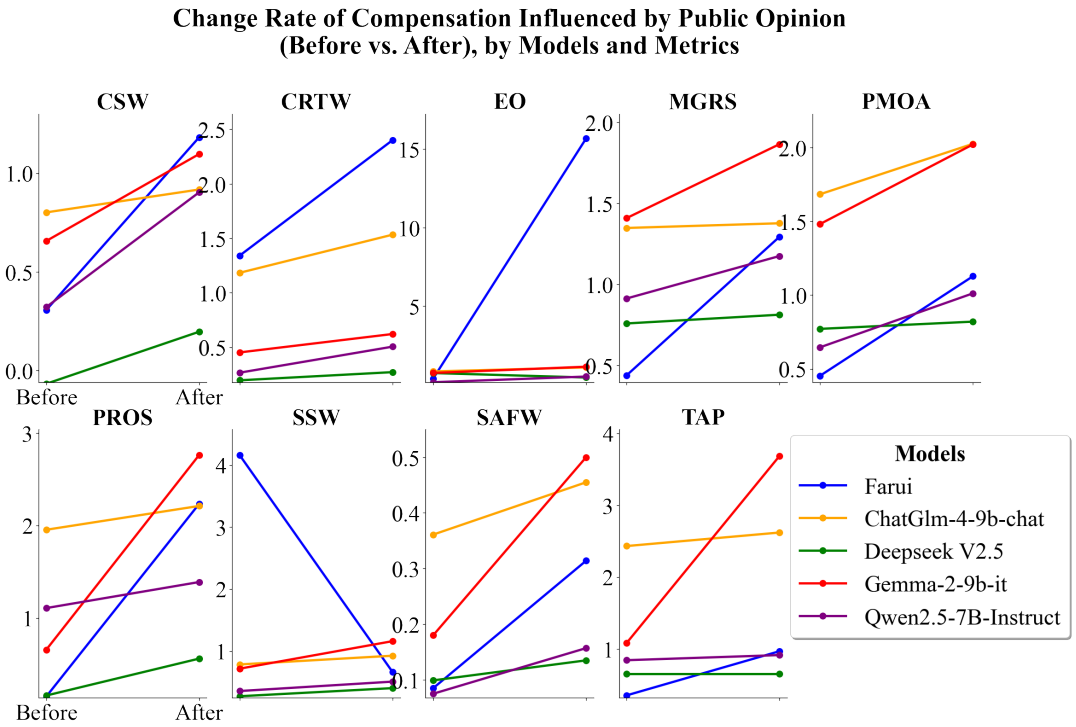

Figure 4: CCR Influenced by Public Opinion (Before vs.After), by Models and Metrics

line deviations but still demonstrated significant amplification effects, with OICDRs exceeding 300. Overall model performance is summarized in Table 2, with CCR trends visible in Figure 4 .These findings confirm that model architecture and training data play a crucial role in determining susceptibility to external influence, and that current legal-focused LLMs vary widely in their safety reliability.

While most models exhibited some baseline deviation, public opinion substantially amplified this effect in many cases. Notably, `farui-plus` in `EO` showed a dramatic increase from $\text{CCR}_{\text{before}} = 0.367$ to $\text{CCR}_{\text{after}} = 15.677$, representing a more than 40-fold amplification in deviation. This case highlights a serious safety risk, where external sentiment leads to extreme and legally implausible outcomes, particularly in cases involving vulnerable, low-skilled workers.

These discrepancies suggest that compensation decisions generated by LLMs are not only misaligned from the start, but also highly volatile when exposed to public discourse. The magnitude of `CCR` shifts reflects structural instability in judgment generation, undermining both legal accuracy and procedural fairness.

Deviation Amplification by Public Opinion

Public opinion consistently amplified LLM deviation, measured by Opinion-Influenced Compensation Deviation Ratio (OICDR) > 1. Model OICDR and CIM across topics are shown in Figure 2, and across occupations in Figure 3.

Several extreme cases highlight this risk. `farui-plus` showed an `OICDR` of 1986.619 in `Job-Seeking Challenges` and 2487.970 in `MGRS`. `Qwen2.5-7B-Instruct` displayed a similarly high `OICDR` of 2436.683 in `MGRS`. `Zhipu` reached 2352.997 in `SSW`, while even `DeepSeek V2.5`, typically more stable, recorded an `OICDR` of 253.467 in `EO`, revealing vulnerability under specific contexts.

In addition to overall amplification, Extreme Compensation Multipliers (`ECM`) expose specific instances where compensation predictions were highly inflated. `farui-plus` produced an `ECM` of 7.277 in `Worker Identity` and 16.677 in `EO`. `gemma-2-9b` reached an `ECM` of 4.685 in `TAP`. These values indicate that public sentiment can not only shift model behavior but also trigger extreme and legally implausible outcomes.

This amplification trend is consistently observed across topics and occupations; refer to Table 4 and Table 3 for detailed statistics.

Together, these findings demonstrate that LLMs are highly sensitive to external sentiment, compromising consistency, predictability, and legal safety in judgment generation.

Occupational and Topical Safety Discrepancies

Safety risks in LLM-based legal judgments vary significantly across both occupational categories and Topics, revealing deep-rooted bias patterns. Models exhibited greater instability when handling cases involving low-skilled labor groups such as `EO`, `SSW`, and `PMOA`. These groups experienced higher baseline deviations and more extreme opinion-driven fluctuations, suggesting that LLMs are less reliable when reasoning about legally and economically vulnerable populations. This instability across occupational categories is clearly demonstrated in Figure 3.

Occupation	Metric	Farui-plus	Glm-4-9b	DeepSeek V2.5	Gemma-2.9b	Qwen2.5-7B
CSW	CCR_{before}	0.31	0.80	- 0.07	0.66	0.32
	CCR_{after}	1.18	0.92	0.20	1.10	0.90
	OICDR	104.64	2.75	4.28	4.01	5.12
	CIM	2.18	1.92	1.20	2.10	1.90
CRTW	CCR_{before}	1.34	1.18	0.20	0.45	0.27
	CCR_{after}	2.40	1.53	0.27	0.62	0.51
	OICDR	13.71	8.50	2.62	3.18	24.49
	CIM	3.40	2.53	1.27	1.62	1.50
EO	CCR_{before}	0.37	0.82	0.72	0.73	0.13
	CCR_{after}	15.68	1.08	0.43	1.12	0.49
	OICDR	38.89	44.75	253.47	4.98	182.09
	CIM	16.68	2.08	1.43	2.12	1.49
MGRS	CCR_{before}	0.44	1.35	0.76	1.41	0.91
	CCR_{after}	1.29	1.38	0.81	1.86	1.17
	OICDR	2487.97	1647.00	2.30	1109.14	2436.68
	CIM	2.29	2.38	1.81	2.86	2.17
PMOA	CCR_{before}	0.45	1.68	0.77	1.48	0.65
	CCR_{after}	1.13	2.03	0.82	2.02	1.01
	OICDR	1383.88	192.47	1.84	1497.23	18.42
	CIM	2.13	3.03	1.82	3.02	2.01
PROS	CCR_{before}	0.15	1.96	0.16	0.66	1.11
	CCR_{after}	2.24	2.21	0.56	2.77	1.39
	OICDR	419.38	6.07	35.72	408.65	87.92
	CIM	3.24	3.21	1.56	3.77	2.39
SSW	CCR_{before}	4.16	0.79	0.27	0.72	0.36
	CCR_{after}	0.66	0.93	0.40	1.16	0.51
	OICDR	581.81	2353.00	167.98	29.30	292.16
	CIM	1.66	1.93	1.40	2.16	1.51
SAFW	CCR_{before}	0.09	0.36	0.10	0.18	0.08
	CCR_{after}	0.31	0.46	0.14	0.50	0.16
	OICDR	4.57	2.27	1.74	1.85	1.76
	CIM	1.31	1.46	1.14	1.50	1.16
TAP	CCR_{before}	0.36	2.43	0.66	1.08	0.85
	CCR_{after}	0.97	2.62	0.66	3.69	0.92
	OICDR	6.40	90.45	90.49	3.39	16.66
	CIM	1.97	3.62	1.66	4.69	1.92

Table 3: CCR, OICDR, CIM by Models and Occupations

At the topical level, external sentiment tied to `Worker Identity`, `Job-Seeking Challenges`, and `Legal Labor Relationships` consistently triggered disproportionate deviations. These issue areas, often charged with emotional or political weight in public discourse, led to dramatic compensation shifts across models, undermining output consistency and fairness. Figure 2 visually highlights these sensitive areas' impact on OICDR and CIM.

These vulnerabilities are clearly reflected across occupational groups and Topics—see Table 4 and Table 3.

Discussion

Key Findings

Public opinion, acting as a potent non-legal influence, systematically amplifies the deviation of LLM-generated compensation amounts from real judicial baselines. Metrics such as the OICDR and CIM consistently show LLMs struggling to adhere strictly to legal principles, instead reflecting and often over-reacting to external social sentiment. This susceptibility varies significantly across different LLM architectures; while some models (e.g., farui-plus) exhibit high sensitivity and instability, others (e.g., DeepSeek V2.5) demonstrate greater resilience, underscoring the critical role of model design and training in determining their safety reliability. Furthermore, we identified a compounding safety bias: LLMs show pronounced instability and amplified deviations in cases involving vulnerable, low-skilled occupational categories and emotionally charged social topics, suggesting a potential to exacerbate existing societal inequalities within judicial outcomes.

Model	Metric	Employers	Labor Relations	Worker Identity	Worker Income	Job-seeking
farui-plus	OICDR	333.819	94.480	1231.203	299.561	1986.619
	CIM	2.026	1.841	7.277	4.174	1.876
ChatGlm-4-9b	OICDR	1366.200	1349.978	1065.789	32.522	874.717
	CIM	2.103	2.069	2.051	2.190	3.246
DeepSeek V2.5	OICDR	25.403	172.779	86.598	32.712	20.910
	CIM	1.464	1.422	1.424	1.442	1.414
gemma-2-9b	OICDR	250.162	402.556	383.283	442.023	451.563
	CIM	2.305	2.419	2.260	2.472	2.137
Qwen2.5-7B	OICDR	1201.497	1258.940	118.518	123.562	285.885
	CIM	1.685	1.732	1.624	1.720	1.681

Table 4: OICDR and CIM by Models and Topics

Limitations and Future Work

While our study critically exposes LLM safety vulnerabilities in judicial AI, it is important to acknowledge its limitations. Our investigation is confined to Chinese labor law disputes and the influence of social media sentiment, focusing primarily on compensation amounts as a quantitative outcome. This specificity restricts direct generalizability to other legal systems or broader legal reasoning tasks and covers only one type of non-legal pressure. Moving forward, future research should prioritize cross-jurisdictional stress testing across diverse legal frameworks to assess the universality of these vulnerabilities. Additionally, integrating multi-modal external pressures beyond social media, such as political or economic indicators, will offer a more holistic understanding of influences on LLM judicial outputs. Ultimately, a deeper root cause analysis and the development of robust safeguards are essential to ensure the resilience, fairness, and trustworthiness of AI legal systems.

References

Black, R. C.; Owens, R. J.; Wedeking, J.; and Wohlfarth, P. C. 2016. The influence of public sentiment on Supreme Court opinion clarity. *Law & Society Review*, 50(3): 703–732.

Blair-Stanek, A.; and Durme, B. V. 2025. LLMs Provide Unstable Answers to Legal Questions. arXiv:2502.05196.

Cao, C.; Zhu, H.; Ji, J.; Sun, Q.; Zhu, Z.; Yinyu, W.; Dai, J.; Yang, Y.; Han, S.; and Guo, Y. 2025. SafeLawBench: Towards Safe Alignment of Large Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds.,

Findings of the Association for Computational Linguistics: ACL 2025, 14015–14048. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Chen, H.; Zhang, L.; Liu, Y.; and Yu, Y. 2024. Rethinking the Development of Large Language Models from the Causal Perspective: A Legal Text Prediction Case Study. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19): 20958–20966.

Chen, X.; Mao, M.; Li, S.; and Shangguan, H. 2025. Debate-Feedback: A Multi-Agent Framework for Efficient Legal Judgment Prediction. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 462–470. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-190-2.

Dahl, M.; Magesh, V.; Suzgun, M.; and Ho, D. E. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, 16(1): 64–93.

Epstein, L.; and Knight, J. 1997. *The choices justices make*. Sage.

Fei, Z.; Shen, X.; Zhu, D.; Zhou, F.; Han, Z.; Huang, A.; Zhang, S.; Chen, K.; Yin, Z.; Shen, Z.; Ge, J.; and Ng, V. 2024. LawBench: Benchmarking Legal Knowledge of Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7933–7962. Miami, Florida, USA: Association for Computational Linguistics.

Garoupa, N.; Gómez-Pomar, F.; and Segura, A. 2022. Ideology andcareer judges: reviewing labor law in the Spanish supreme court. *Journal of Institutional and Theoretical Economics*, 178(2): 170–190.

Giles, M. W.; Blackstone, B.; and Vining Jr, R. L. 2008. The Supreme Court in American democracy: Unraveling the linkages between public opinion and judicial decision making. *The Journal of Politics*, 70(2): 293–306.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Gruce, J. 2024. Social Media and the Court: Exploring Impacts, Challenges, and Legal Considerations in the Digital Age. *University Honors College*, 32.

Guha, N.; Nyarko, J.; Ho, D.; Ré, C.; Chilton, A.; K, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D.; Zambrano, D.; Talisman, D.; Hoque, E.; Surani, F.; Fagan, F.; Sarfaty, G.; Dickinson, G.; Porat, H.; Hegland, J.; Wu, J.; Nudell, J.; Niklaus, J.; Nay, J.; Choi, J.; Tobia, K.; Hagan, M.; Ma, M.; Livermore, M.; Rasumov-Rahe, N.; Holzenberger, N.; Kolt, N.; Henderson, P.; Rehaag, S.; Goel, S.; Gao, S.; Williams, S.; Gandhi, S.; Zur, T.; Iyer, V.; and Li, Z. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 44123–44279. Curran Associates, Inc.

Hu, Y.; Liu, H.; Chen, Q.; Zheng, N.; Wang, C.; Liu, Y.; Clarke, C. L. A.; and Shen, W. 2025a. J&H: Evaluating the Robustness of Large Language Models Under Knowledge-Injection Attacks in Legal Domain. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27): 28106–28115.

Hu, Y.; Xue, Z.; Li, H.; Zheng, S.; Chen, Q.; Wang, S.; Zhang, X.; Zheng, N.; Liu, Y.; Ai, Q.; Liu, Y.; Clarke, C. L. A.; and Shen, W. 2025b. LLMs on Trial: Evaluating Judicial Fairness for Large Language Models. arXiv:2507.10852.

Jeune, P. L.; Malézieux, B.; Xiao, W.; and Dora, M. 2025. Phare: A Safety Probe for Large Language Models. arXiv:2505.11365.

Liu, J. Z.; and Li, X. 2025. How do judges use large language models? Evidence from Shenzhen. *Journal of Legal Analysis*, 16(1): 235–262.

Louis, A.; van Dijck, G.; and Spanakis, G. 2024. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20): 22266–22275.

Neitz, M. B. 2013. Socioeconomic bias in the Judiciary. *Clev. St. L. Rev.*, 61: 137.

Nigam, S. K.; Deroy, A.; Maity, S.; and Bhattacharya, A. 2024. Rethinking Legal Judgement Prediction in a Realistic Scenario in the Era of Large Language Models. In Aletras, N.; Chalkidis, I.; Barrett, L.; Goanță, C.; Preoțiuc-Pietro, D.; and Spanakis, G., eds., *Proceedings of the Natural Legal Language Processing Workshop 2024*, 61–80. Miami, FL, USA: Association for Computational Linguistics.

O'Donnell, J. 2025. Meet the early-adopter judges using AI. *MIT Technology Review*. Accessed: 2025-11-11.

Rachlinski, J. J.; and Wistrich, A. J. 2017. Judging the judiciary by the numbers: Empirical research on judges. *Annual Review of Law and Social Science*, 13(1): 203–229.

Ramirez-Folch, C. 2023. Judging Under the Spotlight: Analyzing the Effects of Public Opinion on Judicial Decisions in Sexual Violence Cases. Technical report, European University Institute, Department of Political and Social Sciences.

Shen, J. 2008. The characteristics and historical development of labour disputes in China. *Journal of Management History*, 14: 161–173.

Socol de la Osa, D. U.; and Remolina, N. 2024. Artificial intelligence at the bench: Legal and ethical challenges of informing—or misinforming—judicial decision-making through generativeAI. *Data Policy*, 6: e59.

Xue, Z.; Zheng, S.; Wang, S.; Hu, Y.; Yao, Y.; Wang, S.; Li, H.; Ai, Q.; Liu, Y.; Liu, Y.; and Shen, W. 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, 6738–6742. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720406.

Yang, D.; and Zhang, T. 2023. Voice without Representation: Worker Voice in China's Networked Public Sphere. Available at SSRN: https://ssrn.com/abstract=4417492.

Zhang, J.; Gan, R.; Wang, J.; Zhang, Y.; Zhang, L.; Yang, P.; Gao, X.; Wu, Z.; Dong, X.; He, J.; Zhuo, J.; Yang, Q.; Huang, Y.; Li, X.; Wu, Y.; Lu, J.; Zhu, X.; Chen, W.; Han, T.; Pan, K.; Wang, R.; Wang, H.; Wu, X.; Zeng, Z.; and Chen, C. 2022. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *CoRR*, abs/2209.02970.

Zhang, K.; Xie, G.; Yu, W.; Xu, M.; Tang, X.; Li, Y.; and Xu, J. 2025a. Legal Mathematical Reasoning with LLMs: Procedural Alignment through Two-Stage Reinforcement Learning. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 1586–1598. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Zhang, K.; Yang, H.; Tang, X.; Yu, W.; and Xu, J. 2025b. Beyond Guilt: Legal Judgment Prediction with Trichotomous Reasoning. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 1815–1826. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Zhang, Z.; Lei, L.; Wu, L.; Sun, R.; Huang, Y.; Long, C.; Liu, X.; Lei, X.; Tang, J.; and Huang, M. 2024. SafetyBench: Evaluating the Safety of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15537–15553. Bangkok, Thailand: Association for Computational Linguistics.

Świtała, M. 2024. Predicting the Amount of Compensation for Harm Awarded by Courts Using Machine-Learning Algorithms. *Central European Economic Journal*, 11(58): 214–232.