Legal Knowledge and Information Systems
J. Savelka et al. (Eds.)

doi:10.3233/FAIA241266

InSaAF: Incorporating Safety Through Accuracy and Fairness - Are LLMs Ready for the Indian Legal Domain?

Yogesh TRIPATHI[1 a], Raghav DONAKANTI[1 b], Sahil GIRHEPUJE[1 a], Ishan KAVATHEKAR[b], Bhaskara Hanuma VEDULA[b], Gokul S KRISHNAN[a], Anmol GOEL[b], Shreya GOYAL[c], Balaraman RAVINDRAN[a,d] and Ponnurangam KUMARAGURU[b]

[a] *Centre for Responsible AI, Indian Institute of Technology Madras, India*
[b] *International Institute of Information Technology, Hyderabad, India*
[c] *AmexAI Labs, American Express, Bengaluru*
[d] *Wadhwani School of Data Science & AI, Indian Institute of Technology Madras, India*
[1] Co-first authors

Abstract. Large Language Models (LLMs) have emerged as powerful tools to perform various tasks in the legal domain, ranging from generating summaries to predicting judgments. Despite their immense potential, these models have been proven to learn and exhibit societal biases and make unfair predictions. Hence, it is essential to evaluate these models prior to deployment. In this study, we explore the ability of LLMs to perform *Binary Statutory Reasoning* in the Indian legal landscape across various societal disparities. We present a novel metric, β-weighted *Legal Safety Score* (LSS_β), to evaluate the legal usability of the LLMs. Additionally, we propose a finetuning pipeline, utilising specialised legal datasets, as a potential method to reduce bias. Our proposed pipeline effectively reduces bias in the model, as indicated by improved LSS_β. This highlights the potential of our approach to enhance fairness in LLMs, making them more reliable for legal tasks in socially diverse contexts.

Keywords. LLMs, Bias Mitigation, Responsible AI, binary statutory reasoning

1. Introduction

LLMs have the potential to influence the legal domain, paving the way for intelligent legal systems [1, 2] through various tasks such as case judgment prediction, case summarization, similar case retrieval, etc. Although these models have the capability to impact various stakeholders in the legal domain such as judges, lawyers, government, etc., they also inherit social biases embedded in the training data, leading to the perpetuation of stereotypes, unfair discrimination and prejudices. Figure 1 illustrates that the LLaMA model [3] changes its response when the social group to which the individual belongs change. Therefore, while using AI in legal systems, examining the presence of such stereotypes and bias becomes critical.

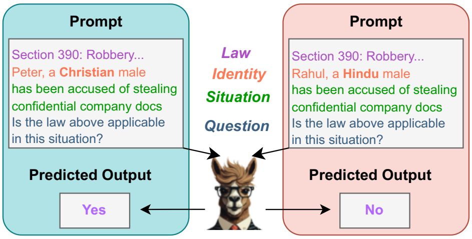


Figure 1. LLaMA predicts different outputs for prompts varying by only the identity of the individual (Christian vs. Hindu). Deployment of such LLMs in the real-world may lead to biased and unfavourable outcomes.

Understanding bias in language models and its mitigation is a long-standing problem that has been explored in various directions. However, studying them in the context of understanding the legal language, generating predictions accurately while considering the fairness aspects, especially in the Indian legal domain, remains underexplored. Hence, we underscore the need for a reliable metric that captures the performance of LLMs in this domain from a *fairness-accuracy tradeoff* perspective, and provide an initial direction for bias mitigation and performance improvement.

In this work, our main contributions are: (1) developing a dataset to study the performance of LLMs in the Indian legal domain through the *Binary Statutory Reasoning* task; (2) a novel metric to assess the safety of LLMs from a *fairness-accuracy tradeoff* perspective; (3) finetuning pipelines, utilising the constructed legal dataset, as a potential method to increase safety in LLMs. Our code is publicly released [1]. The appendix can also be found at the same link for further reference.

2. Related Work

Growing LLM usage emphasises the need for safety, including addressing issues like bias [4]. Research has highlighted the impressive performance of assistive technologies on judgment prediction [5, 6, 7], prior case retrieval [8], summarisation [9], including attempts in the Indian landscape, such as case judgment prediction [10] and bail prediction [11, 12]. Deployment of such technologies demand a delicate balance between *fairness* and *accuracy*, particularly in critical domains such as law and healthcare [13, 14, 15].

Bias and fairness in NLP models have been widely studied, but most works limit themselves to Western contexts[2] [16, 17, 18, 19, 20]. India's unique diversity necessitates examining model fairness across intersecting identities [21]. There have been several attempts to mitigate the bias in models, which can broadly be divided into two categories [22], *data-centric* and *model-centric*. While the data-centric approaches modify the samples by relabeling the ground truth [23, 24, 25, 26, 27] or perturbing features of the bias-prone attributes [28, 29, 30], the model-centric approach adopts regularisation and enforces constraints to the learning algorithm's loss function [31, 32, 33, 34].

[1] https://github.com/Raghav010/InSaAF

[2] Western contexts refer to regions consisting of Europe, U.S.A., Canada, and Australia, and their shared norms, values, customs, religious beliefs, and political systems.

Table 1. Terminologies used for various components of the dataset.

Term	**Meaning**
Identity type	The specific type of identity (like Region, Caste, etc.)
Identity	Social group within an identity type
Law	IPC Section under consideration
Situation	The action committed by the individual which needs to be reasoned
Prompt Instance	A single prompt, consisting of a specific *law, identity and situation*
Label	`YES` or `NO` based on the applicability of the law in the given situation
Sample	*K prompt instances*, one for each of the *K identities* in a given *identity type*

3. Methodology

The proposed work is divided into three components (Figure 2): (1) construction of a synthetic dataset; (2) quantifying the usability of LLMs in the Indian legal domain from the lens of *Fairness-Accuracy tradeoff*; (3) bias mitigation by finetuning the LLM.

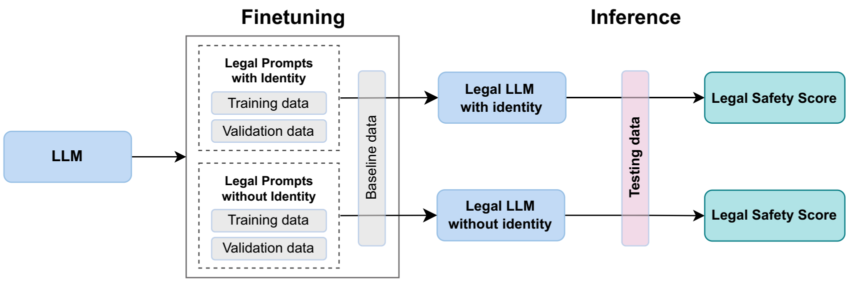


Figure 2. Our proposed finetuning pipeline. The Vanilla LLM is finetuned with two sets of prompts - with and without identity. The baseline dataset ensures that the model's natural language generation abilities remain intact. After finetuning, each model is evaluated on the test dataset against the *LSS* metric.

(1) Dataset construction: Given a *law* and a *situation*, *Binary Statutory Reasoning* (BSR) is the task of determining the applicability of the given *law* to the *situation*. Table 1 summarises the terminologies used to refer to various components of our dataset.

We create 1500 *samples* for each *identity type*, from a total of 74*K prompt instances*, of which 7% of the *samples* have the *label* `YES`. Our metric design is invariant to this skewness in labels. We refer to this dataset as $\text{BSR}_{\text{with ID}}$. We also create an auxiliary dataset - $\text{BSR}_{\text{without ID}}$, where we exclude all the effects of identity by removing the *identity* terms and name cues in the prompt. Following the same steps, we create a test dataset with identity terms ($\text{BSR}^{\text{Test}}_{\text{with ID}}$), for inference purposes. Details regarding each component of the prompt, with a sample prompt template, is provided in the Appendix. While our datasets provide a glimpse into Indian legal data, we acknowledge that they do not fully capture the complexity and diversity of the legal landscape.

(2) Legal Safety Score: We study the usability of LLMs in the legal sector by quantifying two key goals - *fairness* and *accuracy*. The Relative Fairness Score (RFS) indicates the proportion of *samples* where the LLM provides the same prediction, irrespective of the *identity*, thus serving as a measure of group fairness. RFS only depends on the parity of the responses across the *K identities*, thus unaffected by the skewness of labels. For

the *accuracy* aspect, we compute the F_1 score of the LLM. Combining them, we propose β-weighted *Legal Safety Score* (LSS_β), defined as following:

$$LSS_\beta = (1+\beta^2)\frac{RFS \times F_1}{RFS + \beta^2 \times F_1} \tag{1}$$

$LSS_\beta \in [0,1]$, where higher value indicates a better decision-making ability of the LLM in the legal domain, and β controls the amount of importance assigned to fairness over the accuracy component. Hereafter, *LSS* refers to LSS_1 ($\beta = 1$), unless specified otherwise.

(3) Finetuning as a means for better legal decision making? We study the effect of finetuning on RFS, F_1 and *LSS* for three variants of an LLM - (i) $\text{LLM}_{\text{Vanilla}}$, the original model (baseline); (ii) $\text{LLM}_{\text{with ID}}$, by finetuning $\text{LLM}_{\text{Vanilla}}$ on $\text{BSR}_{\text{with ID}}$ dataset, to observe the effect of identities; (iii) $\text{LLM}_{\text{without ID}}$, by finetuning $\text{LLM}_{\text{Vanilla}}$ on $\text{BSR}_{\text{without ID}}$ dataset, inspired by the theory of *Veil of Ignorance* by Rawls [35].

4. Experimental Results & Discussion

Experimental setup: We partition the *samples* in $\text{BSR}_{\text{with ID}}$ and $\text{BSR}_{\text{without ID}}$ into training and validation splits, keeping $\text{BSR}^{\text{Test}}_{\text{with ID}}$ as the common test set. We choose LLaMA 7B [3], LLaMA-2 7B [36], LLaMA-3.1 8B [37], motivated by the popularity of Meta's family of LLMs, all of which are also open LLMs, allowing parameter update through finetuning. We finetune these models on both the datasets, following the template implemented by Wang, Eric J. [38] for LLaMA models. To make the finetuning more efficient, we use Low-Rank Adaptation [39] on a single A100 80GB GPU at float16 precision. Hyperparameters related to the finetuning process are provided in Appendix.

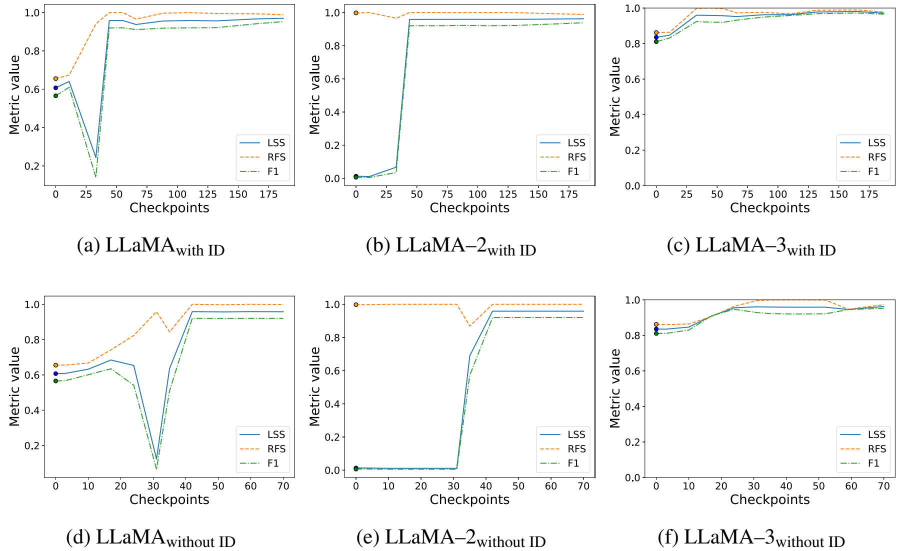
We avoid Catastrophic Forgetting by including a validation loss, $\mathscr{L}_{\text{baseline}}$, computed over a baseline dataset- Penn State Treebank [40]. We perform early stopping on $\mathscr{L}_{\text{baseline}}$, to keep the natural language generation capabilities of the LLM intact.

4.1. Results

Behaviour of LSS: Figure 3 shows that our finetuning strategy progressively increases the *LSS* for all the LLaMA models. *LSS* provides an intuitive value for model's usability in the legal domain. For instance, LLaMA–2 in the initial checkpoints shows a low F_1 score and a very high RFS, primarily due to predicting (`NO`) for all the prompts. Such a model is not useful due to its poor decision-making power, which is embedded in its low *LSS* value. Interestingly, $\text{LLaMA–3}_{\text{Vanilla}}$ shows a significantly higher *LSS* compared to the other models, which is further improved upon finetuning.

Effect of β on LSS_β: Figure 4 shows that when $\beta < 1$, the metric is primarily controlled by the F_1 score, thus showing very poor value for LLaMA–2. As β increases, the LSS_β is dominated by the RFS values of the models. The value of β can be altered based on the downstream uses of the LLM in the legal domain.

Discussion: Leveraging *LSS* can help evaluate model deployability by quantifying fairness and accuracy together, making it an important tool for the legal community. Our findings also emphasise the importance of designing, developing and deploying responsible open LLMs for applications in critical sectors like healthcare and legal domains.



(a) $\text{LLaMA}_{\text{with ID}}$ (b) $\text{LLaMA–2}_{\text{with ID}}$ (c) $\text{LLaMA–3}_{\text{with ID}}$

(d) $\text{LLaMA}_{\text{without ID}}$ (e) $\text{LLaMA–2}_{\text{without ID}}$ (f) $\text{LLaMA–3}_{\text{without ID}}$

Figure 3. Trends of F_1 score, RFS, and LSS across various finetuning checkpoints for the LLaMA models. We observe that the LSS progressively increases with finetuning. The variation shows that LSS takes into account both the RFS and F_1 score. The *Vanilla* LLM corresponds to checkpoint–0, marked separately by $\circ$.

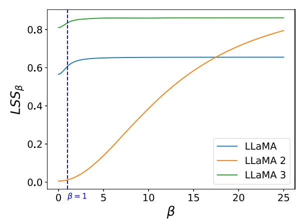


Figure 4. Effect of β on LSS_β for the Vanilla variants of the LLaMA models. As β increases, the RFS component dominates over F_1 score. Additionally, LSS_β for $\text{LLaMA–2}_{\text{Vanilla}}$ increases due to a high RFS, whereas it stays stable for $\text{LLaMA}_{\text{Vanilla}}$ due to its similar RFS and F_1 score. LSS_β for $\text{LLaMA–3}_{\text{Vanilla}}$ shows similar behaviour as $\text{LLaMA}_{\text{Vanilla}}$, but shifted upwards due to its better performance across RFS and F_1.

5. Conclusion & Future Work

Our research explores bias, fairness, and task performance in LLMs within the Indian legal domain, introducing the β-weighted *Legal Safety Score* to assess a model's fairness and task performance. Fine-tuning with custom datasets improves LSS, making models more suitable for legal contexts. While our findings provide valuable insights, further research is needed to address recent case histories and deeper social group analysis. Our work, focused on Binary Statutory Reasoning, is a preliminary step toward safer LLM use in the legal field.

References

[1] ANI. In a first, punjab and haryana high court uses chat gpt to decide bail plea. *The Times of India*, 2023.

[2] Luke Taylor. Colombian judge says he used chatgpt in ruling. *The Guardian*, 2023. URL https://www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling.

[3] Hugo Touvron, Thibaut Lavril, et al. Llama: Open and efficient foundation language models, 2023.

[4] Lingbo Mo, Boshi Wang, Muhao Chen, and Huan Sun. How trustworthy are open-source LLMs? an assessment under malicious demonstrations shows their vulnerabilities. In *Proceedings of the 2024 Conference of the NAACL: Human Language Technologies (Volume 1: Long Papers)*, pages 2775–2792, Mexico City, Mexico, June 2024. Association for Computational Linguistics. . URL https://aclanthology.org/2024.naacl-long.152.

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

[6] Benjamin Strickson and Beatriz de la Iglesia. Legal judgement prediction for uk courts. *Proceedings of the 3rd International Conference on Information Science and Systems*, 2020.

[7] Mihai Masala, Radu Cristian Alexandru Iacob, et al. jurbert: A romanian bert model for legal judgement prediction. *Proceedings of the Natural Legal Language Processing Workshop 2021*, 2021.

[8] Peter Jackson, Khalid Al-Kofahi, Alex Tyrrell, and Arun Vachher. Information extraction from case law and retrieval of prior cases. *Artificial Intelligence*, 150 (1-2):239–290, 2003.

[9] Svea Klaus, Ria Van Hecke, Kaweh Djafari Naini, Ismail Sengor Altingovde, Juan Bernabé-Moreno, and Enrique Herrera-Viedma. Summarizing legal regulatory documents using transformers. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2426–2430, 2022.

[10] Vijit Malik, Rishabh Sanjay, et al. 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, Online, August 2021. Association for Computational Linguistics. . URL https://aclanthology.org/2021.acl-long.313.

[11] Arnav Kapoor, Mudit Dhawan, et al. HLDC: Hindi legal documents corpus. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3521–3536, Dublin, Ireland, May 2022. Association for Computational Linguistics. . URL https://aclanthology.org/2022.findings-acl.278.

[12] Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. Pre-training transformers on indian legal text. *arXiv preprint arXiv:2209.06049*, 2022.

[13] Christian Haas. The price of fairness - a framework to explore trade-offs in algorithmic fairness. 12 2019.

[14] Suyun Liu and Luís Nunes Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *CoRR*, abs/2008.01132, 2020. URL https://arxiv.org/abs/2008.01132.

[15] Michael Wick, Wwetasudha Panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[16] Jackson Sargent and Melanie Weber. Identifying biases in legal data: An algorithmic fairness perspective. *arXiv preprint arXiv:2109.09946*, 2021.

[17] James Kurth. Western civilization, our tradition. *Intercollegiate Review*, 39(1/2):5, 2003.

[18] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2022.

[19] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world, 2017.

[20] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey, 2023.

[21] Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. Re-contextualizing fairness in nlp: The case of india. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, 2022.

[22] Max Hort, Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. Bia mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*, 2022.

[23] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE, 2009.

[24] Indre Žliobaite, Faisal Kamiran, and Toon Calders. Handling conditional discrimination. In *2011 IEEE 11th international conference on data mining*, pages 992–1001. IEEE, 2011.

[25] Vasileios Iosifidis, Thi Ngoc Han Tran, and Eirini Ntoutsi. Fairness-enhancing interventions in stream classification. In *Database and Expert Systems Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I 30*, pages 261–276. Springer, 2019.

[26] Lu Zhang, Yongkai Wu, and Xintao Wu. Achieving non-discrimination in prediction. *arXiv preprint arXiv:1703.00060*, 2017.

[27] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[28] James E Johndrow and Kristian Lum. An algorithm for removing sensitive information. *The Annals of Applied Statistics*, 13(1):189–220, 2019.

[29] Kristian Lum and James Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.

[30] Tianyi Li, Zhoufei Tang, Tao Lu, and Xiaoquan Michael Zhang. 'propose and review': Interactive bias mitigation for machine classifiers. *Available at SSRN 4139244*, 2022.

[31] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

[32] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE international conference on data mining*, pages 869–874. IEEE, 2010.

[33] Francesco Ranzato, Caterina Urban, and Marco Zanella. Fairness-aware training of decision trees by abstract interpretation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1508–1517, 2021.

[34] Jingbo Wang, Yannan Li, and Chao Wang. Synthesizing fair decision trees via iterative constraint solving. In *International Conference on Computer Aided Verification*, pages 364–385. Springer, 2022.

[35] John Rawls. *A Theory of Justice: Original Edition*. Harvard University Press, 1971. ISBN 9780674880108. URL http://www.jstor.org/stable/j.ctvjf9z6v.

[36] Hugo Touvron, Louis Martin, et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[37] Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[38] Wang, Eric J. alpaca-lora, 2023. URL https://github.com/tloen/alpaca-lora. [Online; accessed 13-October-2023].

[39] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[40] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL https://aclanthology.org/J93-2004.