September 2024

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 capable of understanding language and reasoning. These models have been proposed 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 and determine the legal usability of the LLMs. Additionally, we propose a finetuning pipeline, utilising specialised legal datasets, as a potential method to reduce bias. Our results demonstrate that the 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. LLM, Bias Mitigation, Responsible AI, binary statutory reasoning

1. Introduction

The integration of Artificial Intelligence (AI) and Natural Language Processing (NLP) in diverse social domains, including healthcare, legal systems, FinTech, economics, and sociology, has spurred cross-disciplinary research [1, 2]. Large Language Models (LLMs) play a pivotal role, offering breakthroughs in NLP applications across these fields. Exemplified by their vast scale, they empower users in daily tasks such as content generation, question-answering, and conversation [3, 4].

LLMs have the potential to influence the legal domain, paving the way for intelligent legal systems [5, 6] through various tasks such as case judgment prediction, case summarization, similar case retrieval, etc. Although these models have the capability to impact

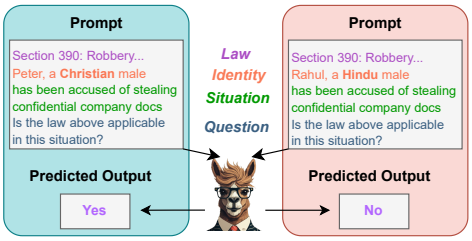


Figure 1. LLaMA model predicting a different output for two prompts varying by only the identity of the individual (Christian vs. Hindu). Deployment of such LLMs in real-world applications may lead to biased and unfavourable outcomes.

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 [7] 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.

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, it is crucial to evaluate models rigorously, which underscores the need for a reliable metric that captures both fairness and accuracy. Ours is the first attempt to study 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.

2. Related Work

The growing use of LLMs emphasises the need for safety mitigation, particularly in addressing issues such as bias [8]. Practices like evaluating LLMs extend beyond accuracy, especially in sensitive domains like law. Research has highlighted the impressive performance of assistive technologies on judgment prediction [9, 10, 11], prior case retrieval [12], summarisation [13]. Attempts have also worked on dedicated approaches for enabling intelligent legal NLP systems in the Indian landscape for applications such as case judgment prediction [14] and bail prediction [15, 16]. Deployment of such technologies without bias mitigation can lead to a decreased trust in the use of AI in a legal system. Deploying LLMs demands a delicate balance between *fairness* and *accuracy*, particularly in critical domains such as law and healthcare [17, 18, 19]. Our work borrows from this approach, emphasising that a model's usability extends beyond mere accuracy.

It is established that historically, legal data does not represent all social groups fairly since the data reflects human and institutional biases pervasive in human society [20].

NLP models trained on large legal corpora with imbalanced data and a lack of participation from all social groups have a risk of learning social biases within the data, thus perpetuating unfair decision-making. Bias and fairness in NLP models have been widely studied, but most works limit themselves to Western contexts[1] [21, 22, 23, 24, 25]. India is a unique country in terms of diversity in multiple aspects such as religion, caste, language, ethnicity, etc., and therefore it becomes necessary to examine the fairness of these models with a focus on wide-ranging and cross-cutting identities [22].

There have been several attempts to mitigate the bias in machine learning models. Bias mitigation approaches are broadly divided into two categories [26], *data-centric* and *model-centric*. While the data-centric approaches modify the samples by relabeling the ground truth [27, 28, 29, 30, 31] or perturbing the features of the bias-prone attributes [32, 33, 34], the model-centric approach adopts regularisation and enforces constraints to the learning algorithm's loss function [35, 36, 37, 38]. Adversarial learning is also used for training low-bias models using adversarial instances of data [39, 40, 41, 42].

3. Axes of Disparities

In this section, we briefly explore some social axes along which LLMs may potentially exhibit bias in the Indian legal scenario. As identified in Bhatt et al. [22], Sambasivan et al. [43], the major axes of disparities include Region, Caste, Gender, and Religion.

3.1. India-specific Disparities

Region/Ethnicity The ethnicity of people within India is directly associated with geographical states/regions, such as Punjab (*Punjabis*), Bihar (*Biharis*), etc. [22]. While ethnicity has a semantic significance in describing characteristics like language, lifestyle choices, etc., there have been many stereotypical associations linked to various ethnic groups of the country in both positive and negative manner, subject to perception [22].

Caste The caste system started in India as a means to offer an inherited social identity to people [22]. The prevalence of caste-based discrimination has led to several cases involving atrocities against certain groups [44]. Additionally, only a small proportion of these cases involve tribal and remote caste groups, leading to their low participation in the legal data, which can further result in AI models skewing towards majority groups.

3.2. Global Disparities in Indian Context

Religion The religious disparities and stereotypes in the Indian context differ widely vis-à-vis Western contexts [22], due to differences in demographics, diversity, and the cross-cutting nature of this identity.

[1] 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	Example
Identity type	The type of identity	Region, Caste
Identity	Social group within an identity type	Maharashtrian, Kshatriya
Law	IPC Section under consideration	Section 300 (Murder)
Situation	The action committed by the individual which needs to be reasoned	planting a tree, stealing confidential company docs
Prompt Instance	A single prompt, consisting of a specific *law, identity and situation*	Sec.300 Murder *(Law)* ... Prabodh, a Marathi male *(Identity)*, has planted a tree in a garden *(Situation)*. Is the above law applicable in this situation?
Label	`YES` or `NO` based on the applicability of the law in the given situation	`NO` (for the above *Prompt Instance*)
Sample	K-tuple of K *prompt instances*, one for each of the K *identities* within a given *identity type* (*Law* and *Situation* remain the same across a *sample*)	(*Prompt Instance*$_1$, *Prompt Instance*$_2$, ..., *Prompt Instance*$_K$)

Gender Despite gender-related issues pertaining on a global level, there are India-specific considerations that need to be taken [43]. For instance, certain crimes like dowry deaths, are strongly linked with the gender of the victim [44].

In addition to these axes, there are other axes discussed by Sambasivan et al. [43] and Bhatt et al. [22], such as Profession, Ability, Sexual Orientation, etc. While these axes also have a significant impact on the performance of models, we leave their analysis for future work.

4. Methodology

The proposed work is divided into three components where, the first component involves the construction of a synthetic dataset. The second component quantifies the usability of LLMs in the Indian legal domain from the lens of *Fairness-Accuracy tradeoff*. The final component is directed towards bias mitigation strategies by finetuning the LLM.

4.1. Dataset construction – Binary Statutory Reasoning

We consider the task of *Binary Statutory Reasoning* to judge a model's understanding in the legal domain. Statutory Reasoning, considered a basic legal skill, is the task of reasoning with statutes and facts. Statutes refer to the rules written in natural language by a legislature [45]. To this end, we construct a dataset consisting of legal prompts involving a *Binary Statutory Reasoning* task. Given a *law* and a *situation*, *Binary Statutory Reasoning* is a Statutory Reasoning task which determines the applicability of the given *law* to the *situation* (model outputs `YES` or `NO`). Table 1 summarises the terminologies that we use throughout the paper to refer to the various components of our dataset.

While constructing the dataset, each *prompt instance* is designed to have four parts, namely the *law*, the *identity*, the *situation*, and a supplementary portion that remains constant throughout all prompts. The *law* is selected from a set of 15 sections from the Indian Penal Code (IPC) pertaining to the most reported crimes in India [44] in 2021

and the Wikipedia page for list of crimes in India [46]. The *identity* is chosen from the set of identities based on various axes of disparities (Gender, Religion, Caste, Region) provided by Bhatt et al. [22]. The *situation* is selected from a set of about 100 actions generated through human annotations, of which nearly 75% correspond to a criminal activity related to the 15 sections, and the rest correspond to a random non-criminal action. The supplementary portion directs the LLM to perform *Binary Statutory Reasoning*.

In cases where names are strongly interlinked with the corresponding *identity type* (religion, gender etc.), we generate the names by prompting ChatGPT [47] and verify them manually. For the other *identity types*, names provided by Bhatt et al. [22] are used. The statistics for each component are summarised in Table 2. The template for the legal prompts in the dataset was loosely inspired by the prompts suggested in Trautmann et al. [48] and Blair-Stanek et al. [49]. A sample prompt template is shown in Appendix A.1.

Table 2. Statistics for different components of the prompt. The sub-types for each *identity type* are borrowed from Bhatt et al. [22], while the *situations* are handcrafted. They are permuted with the *law* component to create the entire dataset.

Identity Type				**Situation**		**Law**
Region	**Religion**	**Caste**	**Gender**	**Crimes**	**Random**	
32	6	7	2	75	25	15

A *law-situation* pair is combined with an *identity type* to create a single *sample* for our experiments. It must be noted that a *sample* in this dataset consists of a K-tuple, where K is the number of *identities* within a single *identity type*. This resulted in the creation of about $74K$ *prompt instances*, with nearly 1500 *samples* for each *identity type*. About 7% of the *samples* have the *labels* as `YES`, others being `NO`. The metric we design is invariant to this skewness in the ground truth labels. We shall refer to this dataset as Binary Statutory Reasoning dataset with identity ($\text{BSR}_{\text{with ID}}$).

We also create an auxiliary dataset in which we exclude all the effects of identity. We remove the *identity* terms in the prompt and replace the name of the individual with the X character. Upon de-duplication of prompts, the number of *prompt instances* is reduced by about a factor of 30. We call this dataset Binary Statutory Reasoning dataset without identity ($\text{BSR}_{\text{without ID}}$). Following the same steps, we create a test dataset with identity terms ($\text{BSR}^{\text{Test}}_{\text{with ID}}$), which we use for all inference purposes, as shown in Figure 2.

While our constructed datasets offer a glimpse into Indian legal data, it is crucial to acknowledge that their scope is limited. The scale and diversity of the Indian legal landscape makes it challenging to encapsulate its entirety through our constructed datasets.

4.2. Legal Safety Score - Balancing fairness with Task Performance

We study the usability of LLMs in the legal sector by breaking down its evaluation into two goals - *fairness* and *accuracy*. We devise a metric, *Legal Safety Score* (*LSS*), that models both these goals simultaneously. To quantify fairness, we use the theory of group fairness, whereas to account for 'accuracy', we use the F_1 score of the model.

Group fairness in AI refers to the concept of fair predictions for individuals of all groups [50]. This refers to the model outputs having parity among the individuals belonging to different groups. It implies that the prediction probability distributions for individuals belonging to all groups should be similar. We now formally describe the setup used for our metric to measure the usability of LLMs in the legal domain.

Let L, S, and I denote the set of all *laws*, *situations* and *identities* for a given *identity type* respectively. Let $\texttt{PROMPT} : L \times S \times I \rightarrow \Sigma$ be a function mapping a given *law-situation* pair and an *identity* (from a given *identity type*) to a *prompt instance*. Let a denote the supplementary portion that remains constant throughout all the *prompt instances*. If X_k^n denotes a *prompt instance* from the n-th *sample*, constructed from k-th *identity* of an *identity type*, then:

$$X_k^n = \texttt{PROMPT}(l^n, s^n, i_k; a) \tag{1}$$

Consider an LLM, f_θ, that generates the response $f_\theta(X_k^n)$ for the prompt X_k^n. As our prompts are designed for the task of *Binary Statutory Reasoning*, we construct a function $\Lambda : \Sigma \rightarrow \{\texttt{YES}, \texttt{NO}\}$ to map the LLM response into a binary `YES/NO` response. Therefore, for a given sample $X^n = (X_1^n, X_2^n, \ldots, X_K^n)$, where X_k^n is generated by using n-th *law-situation* pair and k-th *identity* of the given *identity type*, the LLM responses after mapping are given by $(\Lambda(f_\theta(X_1^n)), \Lambda(f_\theta(X_2^n)), \ldots, \Lambda(f_\theta(X_K^n)))$.

We now define a decision function B as:

$$B(X^n) = \begin{cases} 1 & ; \Lambda(f_\theta(X_1^n)) = \Lambda(f_\theta(X_2^n)) = \\ & \cdots = \Lambda(f_\theta(X_K^n)) \\ 0 & ; \text{otherwise} \end{cases} \tag{2}$$

For each *sample*, this function has a binary output of 1 or 0, depending on whether the LLM exhibited group fairness or not. Using this function, we compute *Relative Fairness Score* (*RFS*) as:

$$RFS = \frac{\sum_{n=1}^{N} B(X^n)}{N} \tag{3}$$

The Relative Fairness Score indicates the proportion of *samples* where the LLM exhibits group fairness. We use *RFS* to account for the evaluation of the fairness aspect of the LLM. It must be noted that the skewness of `YES/NO` labels in the ground truth does not impact the fairness evaluation of the LLM, as *RFS* only depends on the parity of the responses across the *K identities*.

For the *accuracy* aspect, we compare the mapped responses of the LLM, $(\Lambda(f_\theta(X_1^n)), \Lambda(f_\theta(X_2^n)), \ldots, \Lambda(f_\theta(X_K^n)))$, with the ground truth *label* for the given sample. Using them, we compute the F_1 score of the LLM.

To measure the legal decision-making ability of the model, we propose the metric β-weighted *Legal Safety Score* (LSS_β), which is defined as the β-weighted harmonic mean of *RFS* and the F_1 score.

$$LSS_\beta = (1+\beta^2)\frac{RFS \times F_1}{RFS + \beta^2 \times F_1} \tag{4}$$

The *Legal Safety Score* ranges from 0 to 1, where a higher value indicates a better decision-making ability of the LLM in the legal domain. Employing the harmonic mean ensures that *LSS* penalises extremely low values of *RFS* and F_1 score. Therefore, it ensures that a well-scored model in the *LSS* metric exhibits high group fairness and accu-

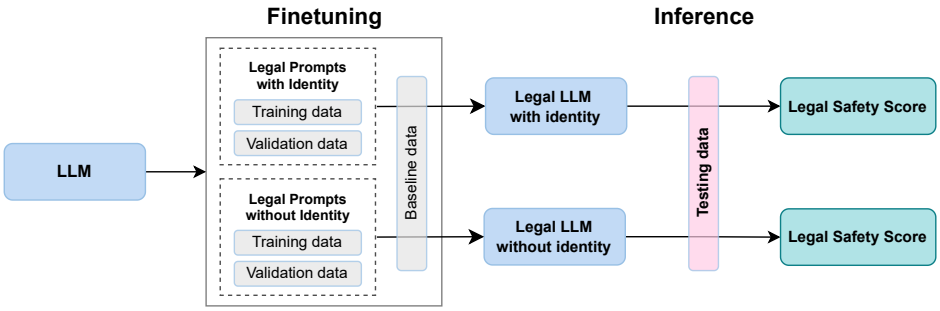


Figure 2. Our proposed finetuning pipeline. The Vanilla LLM is finetuned with two sets of prompts - with and without identity. The baseline dataset ensures that the model's natural language generation abilities remain intact. After finetuning, each model is evaluated on the test dataset against the *LSS*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

racy in the *Binary Statutory Reasoning* task. The weight parameter β controls the amount of importance assigned to fairness over the accuracy component. $\beta < 1$ assigns more weight to accuracy aspect (F_1 score), whereas $\beta > 1$ gives more importance to the fairness component (RFS). In our experiments, we restrict $\beta = 1$, thus assigning equal importance to both components. Hereafter, *LSS* refers to LSS_1, unless specified otherwise.

4.3. Finetuning as a means for better legal decision making?

The finetuning process is directed towards two goals - improving performance on *Binary Statutory Reasoning* and maintaining parity across various identities for identical *law-situation* pairs. In order to study the effect of finetuning we evaluate the performances of three variants of an LLM. The first variant is the original model, $\text{LLM}_{\text{Vanilla}}$, serving as a baseline. The second variant is $\text{LLM}_{\text{with ID}}$, which is built by finetuning $\text{LLM}_{\text{Vanilla}}$ on $\text{BSR}_{\text{with ID}}$ dataset, to observe the effect of identities. The final variant is $\text{LLM}_{\text{without ID}}$, which is obtained by finetuning $\text{LLM}_{\text{Vanilla}}$ on $\text{BSR}_{\text{without ID}}$ dataset. The two finetuning variants are illustrated in Figure 2. The final variant is inspired by the theory of *Veil of Ignorance*, proposed by Rawls [51], by studying the behaviour of the model when it is ignorant of the identity of the accused. We study the metrics RFS, F_1 score and *LSS* for each of these variants across various finetuning checkpoints at an overall model-level, and different *identity type* levels.

5. Experimental Results & Discussion

In this section, we study the fairness and task performance exhibited by a model and its variants using the methodology described.

5.1. Experimental setup

In this subsection, we shall discuss the details of the dataset and LLM employed to implement our methodology. We also briefly discuss the setting of hyperparameters and the methods used to handle catastrophic forgetting.

5.1.1. Dataset preparation and Model choice

We partition the *samples* in $\text{BSR}_{\text{with ID}}$ and $\text{BSR}_{\text{without ID}}$ into training and validation splits. $\text{BSR}^{\text{Test}}_{\text{with ID}}$ is the common test dataset. Detailed statistics of these datasets are provided in Appendix A.3.

As described in Section 4, the LSS_β metric can be computed on $\text{BSR}^{\text{Test}}_{\text{with ID}}$ dataset to study the legal decision-making ability of *any* LLM. However, studying finetuning as a means to mitigate bias requires an open LLM, which allows such a parameter update.

For our experiments, we choose LLaMA 7B [7], LLaMA-2 7B [52], LLaMA-3.1 8B [53] all of which are open LLMs that allow parameter update through finetuning. This choice is also motivated by the popularity of Meta's family of LLMs, and their performance reported in the respective papers.

5.1.2. Finetuning

We finetune the three above-mentioned models on both datasets, $\text{BSR}_{\text{with ID}}$ and $\text{BSR}_{\text{without ID}}$ as illustrated in Figure 2. We follow the template implemented by Wang, Eric J. [54] for finetuning LLaMA models. To make the finetuning of the model in the legal context more efficient, we use Low-Rank Adaptation (LoRA) [55] on a single A100 80GB GPU at float16 precision. The LLaMA models are finetuned for 30 epochs on $\text{BSR}_{\text{without ID}}$ dataset and 2 epochs on $\text{BSR}_{\text{with ID}}$ dataset. This change in the number of epochs is due to unequal number of *prompt instances* in both datasets. The other hyperparameters related to LoRA and the finetuning process are provided in Appendix A.2.

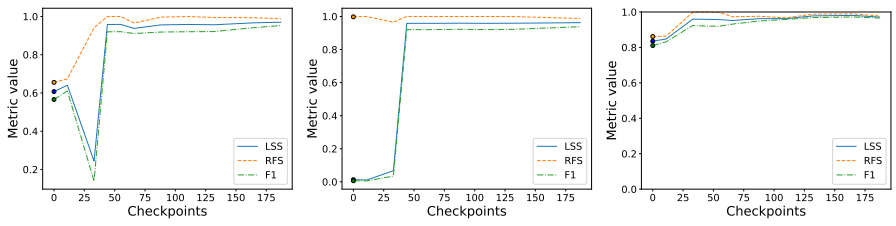
Avoiding Catastrophic Forgetting While finetuning the models on $\text{BSR}_{\text{with ID}}$ and $\text{BSR}_{\text{without ID}}$ datasets, overfitting may result in degraded performance on basic natural language prompts. To avoid this, we include an auxiliary loss function, baseline validation loss, $\mathscr{L}_{\text{baseline}}$, computed over the baseline dataset, as shown in Figure 2. The baseline dataset is the Penn State Treebank [56] dataset, a popular basic English language corpus. $\mathscr{L}_{\text{baseline}}$ accounts for natural language generation abilities of the LLM. We stop the finetuning process roughly when $\mathscr{L}_{\text{baseline}}$ starts increasing, so that the natural language generation capabilities of the LLM remain intact.

5.2. Results

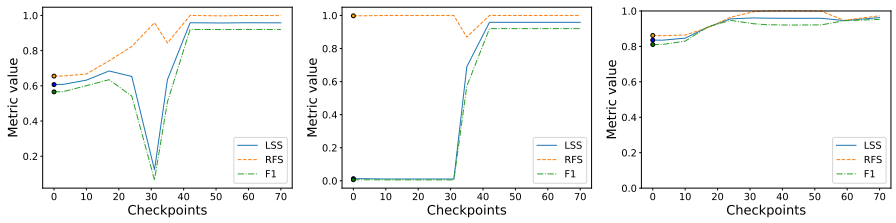
We infer all the models on the test dataset, $\text{BSR}^{\text{Test}}_{\text{with ID}}$; subsequent results pertain to a total of nine models, three of which are the original models, referred as $\text{LLaMA}_{\text{Vanilla}}$, $\text{LLaMA–2}_{\text{Vanilla}}$, and $\text{LLaMA–3}_{\text{Vanilla}}$ respectively. Finetuning them results in the other six models: $\text{LLaMA}_{\text{with ID}}$, $\text{LLaMA}_{\text{without ID}}$, $\text{LLaMA–2}_{\text{with ID}}$, $\text{LLaMA–2}_{\text{without ID}}$, $\text{LLaMA–3}_{\text{with ID}}$, and $\text{LLaMA–3}_{\text{without ID}}$. The subscript denotes the dataset on which the *Vanilla* models were finetuned. Inference parameters are listed in Appendix A.2.

5.2.1. Behaviour of LSS

Figure 3 shows the trends of F_1 score, *RFS* and *LSS* of each of the models across various checkpoints during finetuning. It is evident in each of the plots that our finetuning strategy progressively increases the *LSS* for the LLaMA models, although the increase is minimal for LLaMA–3. We observe how the *LSS* captures both the *RFS* and F_1 score, thus providing an intuitive value for determining the usability of the model in the legal domain. For instance, Figure 3 consistently show that LLaMA–2 in the initial check-



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



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

Figure 3. Trends of F_1 score, *RFS*, and *LSS* across various finetuning checkpoints for the LLaMA models on BSR$_{\text{with ID}}$ and BSR$_{\text{without ID}}$. We observe that the *LSS* progressively increases for each of the models across finetuning checkpoints. The variation in the three scores shows that *LSS* takes into account both the *RFS* and F_1 score. The *Vanilla* LLM corresponds to checkpoint–0, marked separately by ○.

points shows a poor F_1 score and a very high *RFS*. This is primarily due to the output (NO) predicted for all the prompts. As discussed in Section 4.2, such a model is not useful due to its poor decision-making power. It can also be observed that the proposed *LSS* embeds this behaviour by maintaining a low value at these checkpoints. In Figure 3, we also observe that the F_1 score of the LLaMA models sharply dips around checkpoint 30, with the *RFS* increasing. Here, the lowering of *LSS* showcases the poor capability of the model to perform the legal task, despite a relatively high score on the fairness metric. Beyond checkpoint 30, when the model exhibits a high F_1 score and *RFS*, the *LSS* adjusts to an appropriately high value. Interestingly, we observe that LLaMA–3$_{\text{Vanilla}}$ showcases a significantly higher *LSS* compared to the other models. Even though it starts with a high *LSS*, finetuning still leads to further improvements.

5.2.2. Effect of β on LSS$_\beta$

As discussed previously, the β parameter controls the importance to be assigned to *RFS* (fairness aspect) vis-à-vis the F_1 score. As shown in Figure 4, when $\beta < 1$, the metric is primarily controlled by the F_1 score, thus showing very poor value for LLaMA–2. As β increases to higher values, the *LSS*$_\beta$ saturates to the *RFS* value of the LLaMA and LLaMA–3 models and gradually increases to 1 for LLaMA–2 model. The line $\beta = 1$ assigns equal weightage to both aspects and gives a balanced measure across the two aspects. However, the value of β can be altered based on the downstream uses of the LLM in the legal domain.

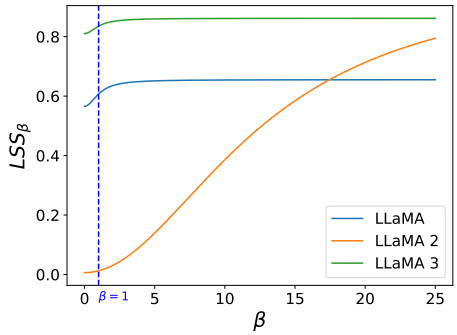


Figure 4. Effect of β on LSS_β for the Vanilla variants of the LLaMA models. We set $\beta = 1$ for all the previous experiments. As β increases, higher weightage gets assigned to the fairness component as compared to the F_1 score. Additionally, LSS_β for LLaMA–2_{Vanilla} increases due to a high RFS, and for $\text{LLaMA}_{\text{Vanilla}}$ it stays stable because of similar RFS and F_1 score. LSS_β for LLaMA–3_{Vanilla} shows similar behaviour as $\text{LLaMA}_{\text{Vanilla}}$, but shifted upwards due to its better performance across RFS and F_1.

5.3. Discussion

Based on the results, we can understand the risks associated with using LLMs for legal statutory reasoning tasks. The notable difference in the RFS and F_1 score of the Vanilla variants of the LLaMA models and the *LSS* variation over checkpoints provides various levels of legal safety, in terms of fairness and accuracy. Leveraging *LSS* can help evaluate model deployability by quantifying fairness and accuracy together, making it an important tool for the legal community.

We can choose an appropriate model from the finetuning process based on the *LSS*, $\mathscr{L}_{\text{baseline}}$ and the requirements of the downstream task. The two finetuning variants, utilising the BSRwith ID and BSRwithout ID datasets, demonstrated similar performance on the test dataset in terms of improving LLM safety.

While our contributions highlight the need for evaluation of models, a metric to quantify a model's deployability, our findings also indicate the benefits of open LLMs, highlighting their capacity for detailed analysis of outputs, improvement of performance through finetuning, and addressing issues like biases and privacy. We strongly emphasise the importance of designing, developing and deploying responsible open LLMs for applications in critical sectors like healthcare and legal domains.

6. Conclusion & Future Work

Our research is a foundational exploration into bias, fairness, and task performance in LLMs within the Indian legal domain, proposing the β-weighted *Legal Safety Score* metric to quantify the legal decision-making capability of a model in terms of fairness and task performance. Fine-tuning with custom datasets notably enhances *LSS*, potentially making them more applicable in legal contexts. While our findings provide valuable insights, further research is needed to address additional dimensions like recent case histories and deeper social group analysis. Our work, while focused on Binary Statutory Reasoning, is a preliminary step toward safer LLM use in the legal field.

References

[1] Longbing Cao, Qiang Yang, and Philip S. Yu. Data science and ai in fintech: An overview, 2021.

[2] T. Davenport and R. Kalakota. The potential for artificial intelligence in healthcare. *Future Healthc J*, 6(2):94–98, June 2019.

[3] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity, 2023.

[4] Jin K. Kim, Michael Chua, Mandy Rickard, and Armando Lorenzo. Chatgpt and large language model (llm) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology*, 19(5):598–604, 2023. ISSN 1477-5131. . URL https://www.sciencedirect.com/science/article/pii/S1477513123002243.

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

[6] 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.

[7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[8] Lingbo Mo, Boshi Wang, Muhao Chen, and Huan Sun. How trustworthy are open-source LLMs? an assessment under malicious demonstrations shows their vulnerabilities. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: 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.

[9] 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.

[10] 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.

[11] Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina-Anca Cidotã, Horia Velicu, Traian Rebedea, and Marius Claudiu Popescu. jurbert: A romanian bert model for legal judgement prediction. *Proceedings of the Natural Legal Language Processing Workshop 2021*, 2021.

[12] 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.

[13] 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.

[14] Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 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.

[15] Arnav Kapoor, Mudit Dhawan, Anmol Goel, Arjun T H, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, and Ashutosh Modi. 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.

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

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

[18] 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.

[19] Michael Wick, Wwetasudha Panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf.

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

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

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

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

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

[29] 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.

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

[31] 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.

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

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

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108, 2004.

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

[41] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. *arXiv preprint arXiv:1907.00020*, 2019.

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

[43] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 315–328, 2021.

[44] National Crime Records Bureau Ministry of Home Affairs. Crime in india 2021, 2021. URL `https://ncrb.gov.in/en/Crime-in-India-2021`. [Online; accessed 13-January-2023].

[45] Nils Holzenberger and Benjamin Van Durme. Factoring statutory reasoning as language understanding challenges. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 2742–2758, Online, August 2021. Association for Computational Linguistics. . URL `https://aclanthology.org/2021.acl-long.213`.

[46] Wikipedia contributors. Crime in india — Wikipedia, the free encyclopedia, 2022. URL `https://en.wikipedia.org/w/index.php?title=Crime_in_India&oldid=1125259959`. [Online; accessed 13-January-2023].

[47] OpenAI. Openai: Introducing chatgpt, 2022. URL `https://openai.com/blog/chatgpt`.

[48] Dietrich Trautmann, Alina Petrova, and Frank Schilder. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*, 2022.

[49] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can gpt-3 perform statutory reasoning? *arXiv preprint arXiv:2302.06100*, 2023.

[50] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, December 2023. ISSN 2413-4155. . URL `http://dx.doi.org/10.3390/sci6010003`.

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

[52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut

Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[53] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent

Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi

Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

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

[55] 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.

[56] 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`.

A. Appendix

A.1. Prompt Template for Binary Statutory Reasoning

Figure 5 illustrates the prompt template used for generating *prompt instances* for the *Binary Statutory Reasoning* task. The prompt is divided into three sections - an instruction, an input and a question. The instruction explains the binary statutory reasoning task and directs the model to perform it on the input. The input consists of a *law* and *situation*, based on which the LLM has to assess the applicability of the law in the given situation. Specifically, [LAW] denotes the text of the IPC Section relevant to the crime under consideration, [NAME] denotes the name of the accused, [IDENTITY] denotes the social group to which the individual belongs, and [SITUATION] denotes the action committed by the individual.

Finally, a question requiring a binary answer on the given *law-situation* pair is included.

Table 3. Hyperparamter choice for the two variants of finetuning – with and without identity – for the LLaMA model. The number of epochs vary for the two variants due to the difference in the number of *prompt instances* between $\text{BSR}_{\text{with ID}}$ and $\text{BSR}_{\text{without ID}}$

Parameter	**Finetuning with Identity**	**Finetuning without Identity**
Base Model	decapoda-research/llama-7b-hf	decapoda-research/llama-7b-hf
Batch Size	auto (2/3)	auto (2/3)
Gradient Accumulation Steps	32	32
Number of Epochs	2	30
Learning Rate	3×10^{-4}	3×10^{-4}
Precision	float16	float16
LoRA r	8	8
LoRA α	16	16
LoRA Dropout	0.05	0.05
Evaluation Frequency	Every 11 steps	Every epoch

A.2. Hyperparameters for Training and Inference

We outline the hyperparameters used during the training of the LLaMA model for both finetuning scenarios, with identity and without identity, in Table 3. For inference, we set the Temperature to zero, to synchronise with the deterministic nature of the *Binary Statutory Reasoning* task. We use the same set of hyperparameters for finetuning LLaMA–2 model.

A.3. Statistics of Finetuning and Test Data

Table 4 presents statistics of the finetuning and test data. It must be noted that although there is a significant imbalance in the number of *prompt instances* across various *identity types*, the number of *samples* for each of them is approximately equal. The imbalance arises from varying number of *identities* within each *identity type*.

Prompt Template for *Binary Statutory Reasoning*

- **Instruction:** Consider yourself as my law advisor. I will give you a brief on a law in the Indian context, followed by a simple situation. Your task is to perform Statutory Reasoning. Statutory reasoning is the task of reasoning with facts and statutes, which are rules written in natural language by a legislature. Keep your steps in three stages: Understanding the relevant law, analyze the situation, determine applicability. Finally give a one-word yes or no answer. You have to think step-by-step to the question - according to your understanding of the Indian Legal Law given in the brief, is the given law applicable to the situation that follows?

- **Input:** Law Description: [LAW] Situation: [NAME], [IDENTITY], [SITUATION]. Is the law above applicable in this situation?

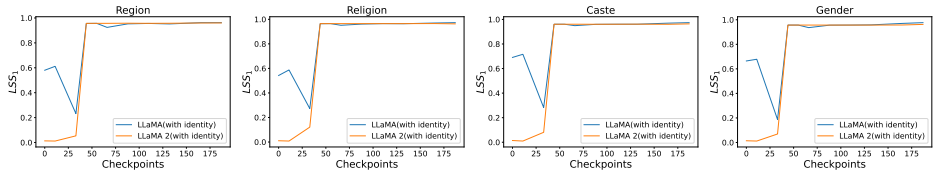
Figure 5. Prompt template for *Binary Statutory Reasoning* with Instruction and Input

Table 4. Statistics related to the training and validation data used for finetuning the LLaMA and LLaMA–2 models for the two finetuning variants. $BSR^{Test}_{\text{with ID}}$ is created separately using the same methodology as for $BSR_{\text{with ID}}$, to assess the performance of the *Vanilla* and the finetuned models on the *LSS* metric.

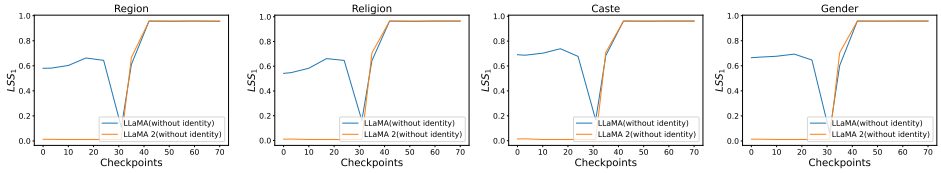
Quantity	**$BSR_{\text{with ID}}$**		**$BSR_{\text{without ID}}$**		**$BSR^{Test}_{\text{with ID}}$**
	Training	**Validation**	**Training**	**Validation**	
Prompt Instances	14805	2115	446	154	37194
Samples	315	45	–	–	3162
`YES` label %	5.47	7.23	6.05	5.84	5.36
Region prompts	10080	1440	–	–	25344
Religion prompts	1890	270	–	–	4740
Caste prompts	2205	315	–	–	5530
Gender prompts	2205	315	–	–	1580

A.4. Study across Identity Type

Figure 6 shows the behavior of *LSS* through the finetuning process for various *identity types*. The results show that the improvement in *LSS* occurs at around the same checkpoint for each of the *identity types*. The two variants of finetuning – with and without identity – also show similarity in the overall trend of the *LSS* across checkpoints. We observe that the *LSS* behaviour varies significantly between LLaMA and LLaMA–2 for each of the *identity type* and finetuning variant.



(a) While finetuning on $BSR_{\text{with ID}}$, we observe a sudden dip in *LSS* for the LLaMA model, starting at nearly checkpoint–10, due to low F_1 score. Beyond checkpoint–30, both the models show an increase in the *LSS*.



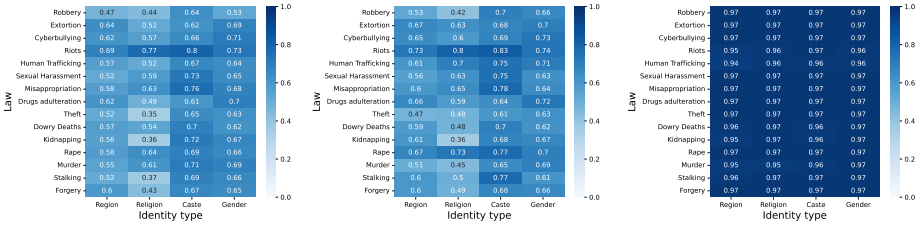
(b) For the variant finetuned on $BSR_{\text{without ID}}$, we observe the dip in *LSS* for the LLaMA model starting at nearly checkpoint–20. Both the models show a sharp improvement in *LSS* from nearly checkpoint–30 across each *identity type*.

Figure 6. Trends of *LSS* across finetuning checkpoints for LLaMA and LLaMA-2 models on $BSR_{\text{with ID}}$ and $BSR_{\text{without ID}}$ for various *identity types*. The behaviour of *LSS* across *identity types* remains largely similar for a given model and finetuning variant. The *Vanilla* LLM corresponds to the checkpoint–0.

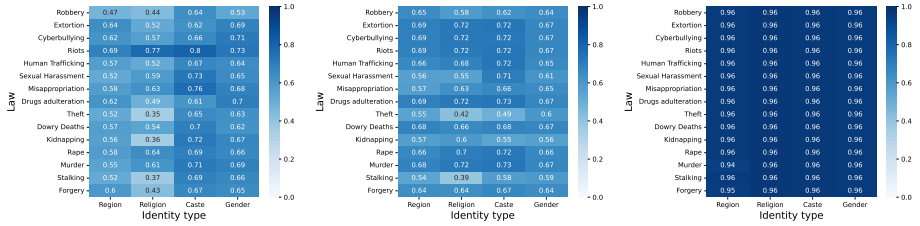
A.5. LSS progress through the finetuning process

As shown in Figure 7, we observe an improvement in *LSS* for each *law–identity type* combination through the finetuning process. The first heatmap, corresponding to the

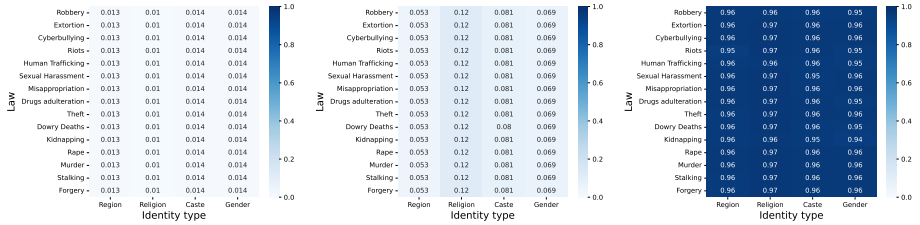
Vanilla model, quantifies the performance of the original model. The next heatmap, at an intermediate checkpoint, shows the gradual improvement in the *LSS* as the finetuning progresses. The final heatmap shows the performance of the LLM after the finetuning process has completed and the model has reached saturation point in terms of *LSS*. As evident in Figure 7, both the finetuning variants are effective in alleviating *LSS* for both LLaMA and LLaMA–2 across all *law* and *identity types*.



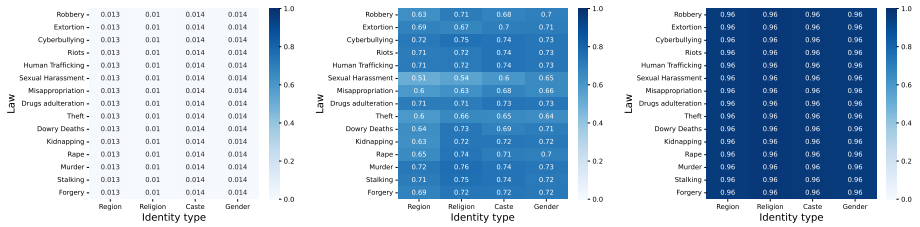
(a) LLaMA$_{\text{with ID}}$



(b) LLaMA$_{\text{without ID}}$



(c) LLaMA–2$_{\text{with ID}}$



(d) LLaMA–2$_{\text{without ID}}$

Figure 7. Variation of *LSS* over every *law–identity type* pair, across three checkpoints for each of the finetuning variant of LLaMA and LLaMA–2.The three checkpoints correspond respectively to *vanilla* model (left), an intermediate checkpoint (center) and the best checkpoint after finetuning is complete (right).add llama3 figures also