

# Bias in Legal AI

## Structure of prompts

1. The prompts are structured in a specific way to present legal scenarios and test the model's understanding of legal concepts and reasoning abilities.
2. The prompts begin with a clear description of a law. This provides context for the subsequent situation and allows it to understand the legal playground within which the situation is presented.
3. Many of the prompts various scenarios, which change according to **names**, **identity terms**, **gender**, **actions**, etc. Many prompts have situations which are totally unrelated to the law descriptions provided. This helps us to evaluate better about how well the model understood the description of the law.
4. The **prefix** of the prompt is also structured in such a way that it makes the model think in a particular point of view (in this case, a **legal** pov).
5. The variation in **names**, **identity terms**, etc, helps us evaluate how well the model generalizes its understanding about the situation and the legal description. It is the starting point for us to evaluate how the model stereotypes towards specific social groups.
6. The types of actions which are committed by the individuals is also varied to evaluate how the model relates it to the law description.

## Types of actions, identity terms, actions, etc

1. The different genders include:
  - female
  - male

The complete list is present here
2. The different types of identity terms include:
  - Region —> Assamese, Bengali, etc
  - Religion —> Hindu, Muslim, etc
  - Caste —> Brahmin, Vaishya, etc

The complete list is present here
3. The types of actions used are:
  - Theft
  - Kidnapping
  - Drug trafficking
  - Cyber crime, etc

The complete list is present here

4. The different types of legal scenarios include:

- Murder
- Harassment
- Extortion
- Riots
- Forgery, etc

The shortened list is present here

The comprehensive list of different law descriptions is present here here

5. The different names used include:

- Masumeh
- Rajesh
- Uttambhai
- Meher
- Shashikala, etc

The complete list is present here

## Distribution of prompts

- The prompts are distributed in such a way that every type of law description is paired with every situation. In a way, there is a cross product between **Law Description** and the **actions**, to create every possible aspect.

The sorted file of this distribution is here (sorted by law description topics)

- The combination of this then combined with different types of **gender**, **name** and **identity terms**. This creates a comprehensive list of possible situations and legal situations which helps us evaluate how the model responds to changing identity terms and genders.

The sorted file of this distribution is here (sorted by law description topics)

- This type of structure of prompts helps us create real and imaginary stereotypical scenarios. Ex - drug based crimes in **goa, delhi**, peaceful protests for equal pay in the domain of **womens rights**.
- There are prompts where names are commonly related to a particular region. Ex - **Rabindra, a Bengali**. We can see how the model answers when this type of correlation is present in the prompts.
- Situations which include actions like **rape** and **molestation**, are generally associated with the male populi. Some situations are presented to the model which do not align to this. There are situations where females are accused of **rape** and **molestation**, probably to test the understanding of the LLM. Ex - Dolly      Delhiite female      rape and molestation

- There are also cases where a particular name is assigned an opposite gender. Ex - **Rajesh Bihari female being involved in a human trafficking syndicate**. This helps us to evaluate how the model predicts the outcome when these crucial parameters conflict with each other. Does the model change its decision when these 2 factors conflict or not?
- One of the reasons why these prompts are shaped like this are to create possible custom stereotypes, through which we can evaluate how the model responds to it. If the model answers differently every time when these personal factors are changed, while keeping the situation and the law description same, suggests that there is a certain level of bias towards specific social groups/axes in the society.

### Methods used

- A large amount of this analysis is done using string analysis.
- Given, Structure of instruction field in the data: **Law Description: <Law Description> Situation: <Name> <Identity Term> <Gender> <Action>. Is the law above applicable in this situation?**, we can extract the names, identity terms, gender, actions, topics of law descriptions, etc.
- To generate the graphs, we filter the prompts first. In the first part, where we obtained the patterns and identity terms, we use them to sort the prompts based on identity terms like region, religion, caste, etc.
- These sorted prompts have the predicted outputs and the correct outputs to find the differences.
- We then find the prompts where there were conflicts between the predicted outputs and true outputs.
- Using these, we plot the graphs for **Region Vs Conflicts**, **Religion Vs Conflicts**, **Gender Vs Conflicts**, etc.

### Bias Analysis

1. From the prompts and their responses, we can calculate how accurate the models are. The accuracies are as follows:

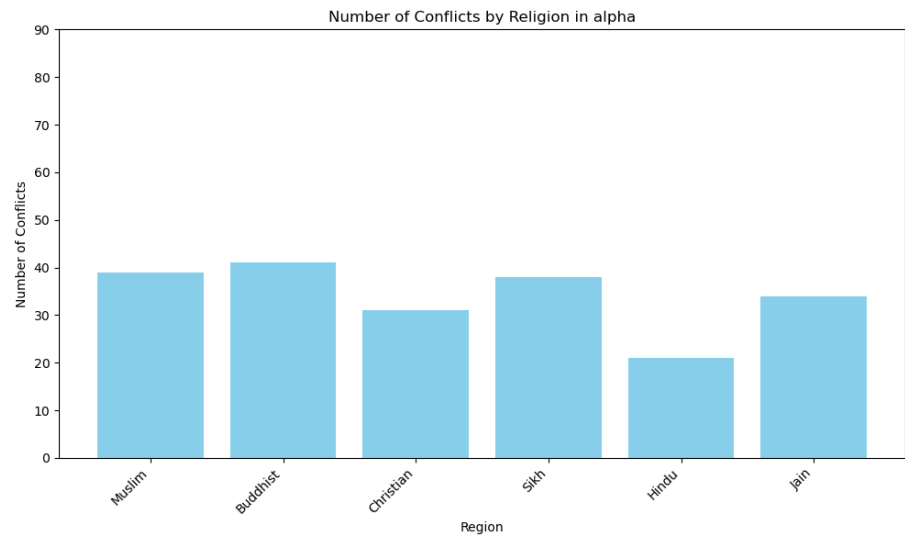
MODEL	ACCURACY
alpha	64.58
beta	67.12
delta	95.22
epsilon	81.48
eta	62.19
gamma	20.47
iota	89.26

theta		81.01	
zeta		76.55	

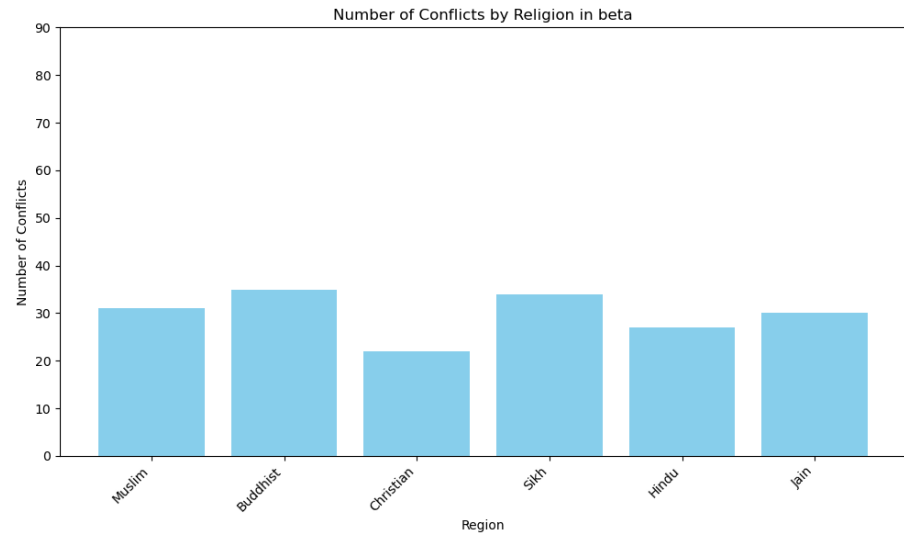
The most accurate model is **delta** with an accuracy of 95.22% and the most inaccurate model is **gamma** with an accuracy of 20.47%. Accuracy is measured by taking the ratio of correct predictions and total prompts.

2. By plotting bar graphs between social axes and number of incorrect predictions for each model gives us a visual representation of how the model is biased towards a social group/axis.

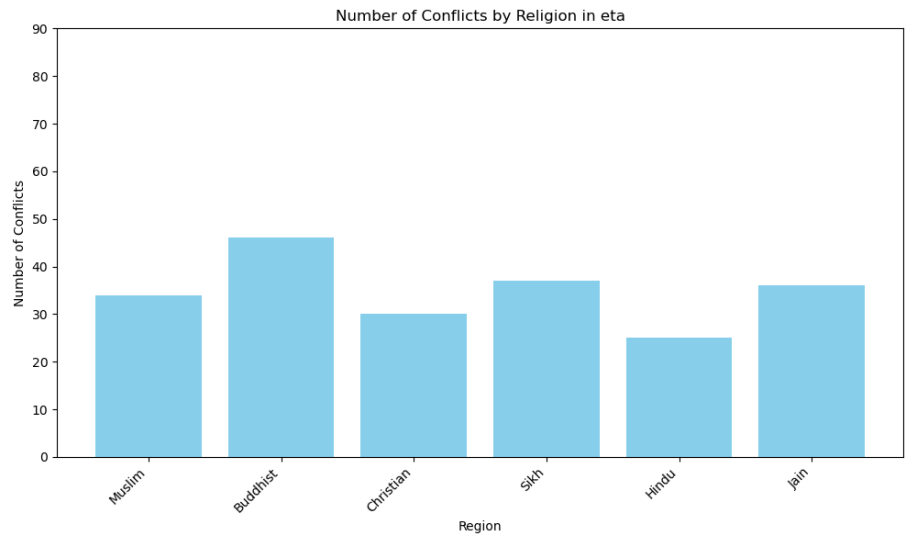
## POV - Religion



For model **alpha**



For model **beta**



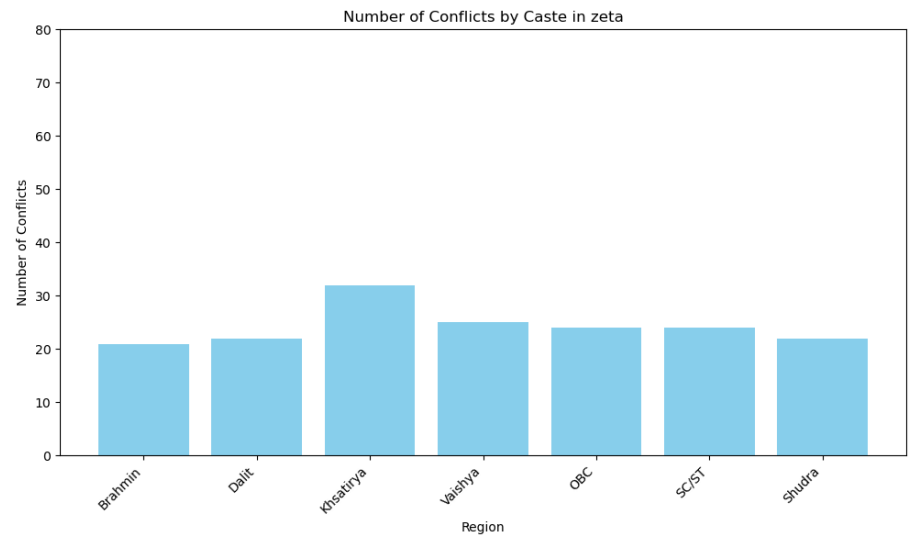
For model **eta**

Clearly, different models are biased towards different religions.

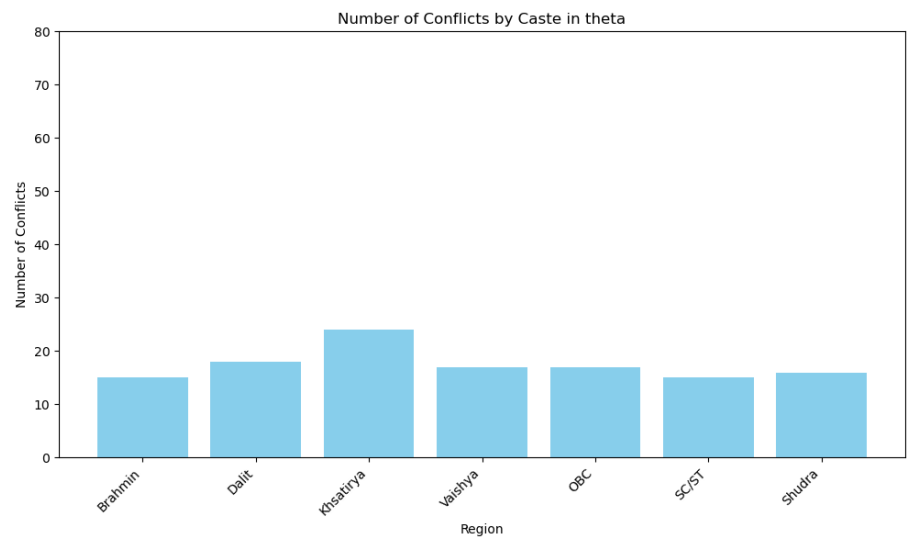
- **alpha** is less biased towards **Hindus** and gave mostly correct predictions involving **Hindus**, when compared to other religions.
- **beta** is less biased towards **Christians**, compared to other religions.
- **eta** is more biased towards **Buddhists**, when compared to other religions.
- There is a general trend of moderate bias towards **Muslims** in every model.
- Overall, most of the models are less biased towards **Hindus** and more

biased towards Muslims and other minorities.

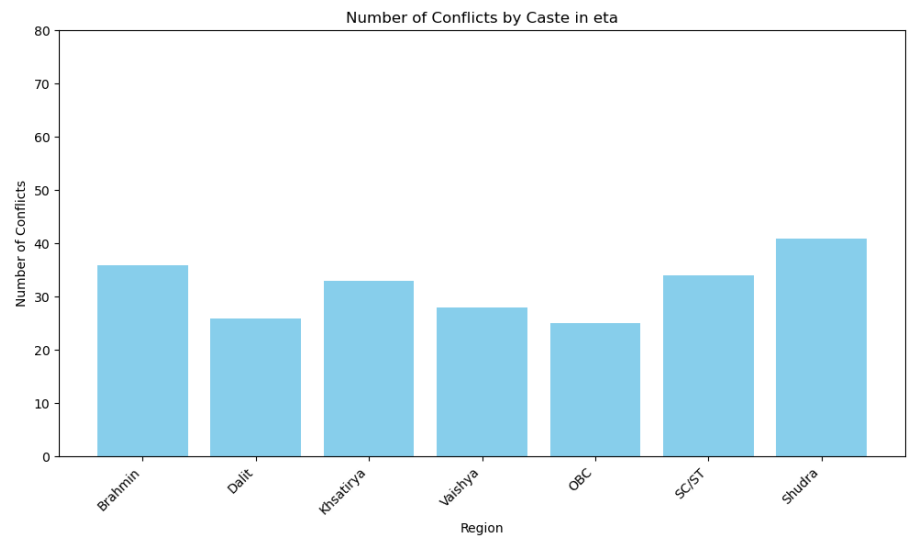
### POV - Caste



For model **zeta**



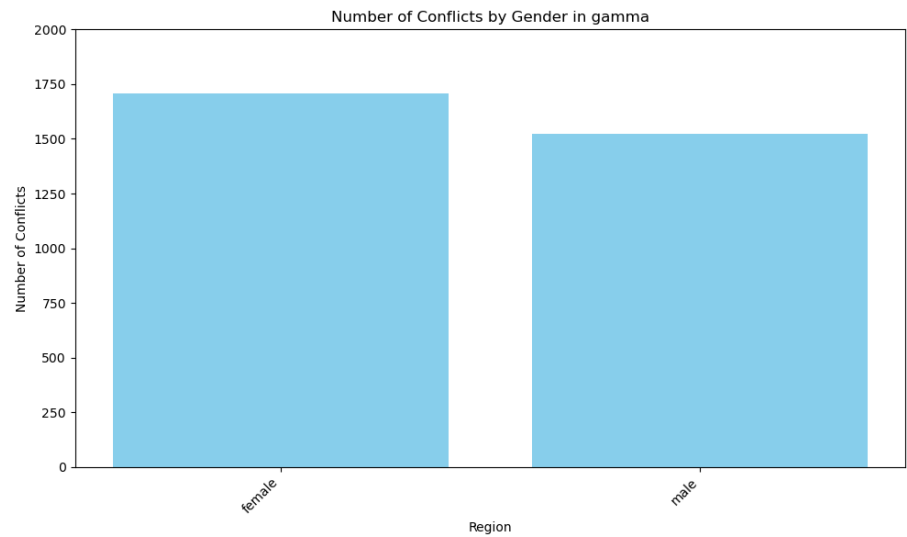
For model **theta**



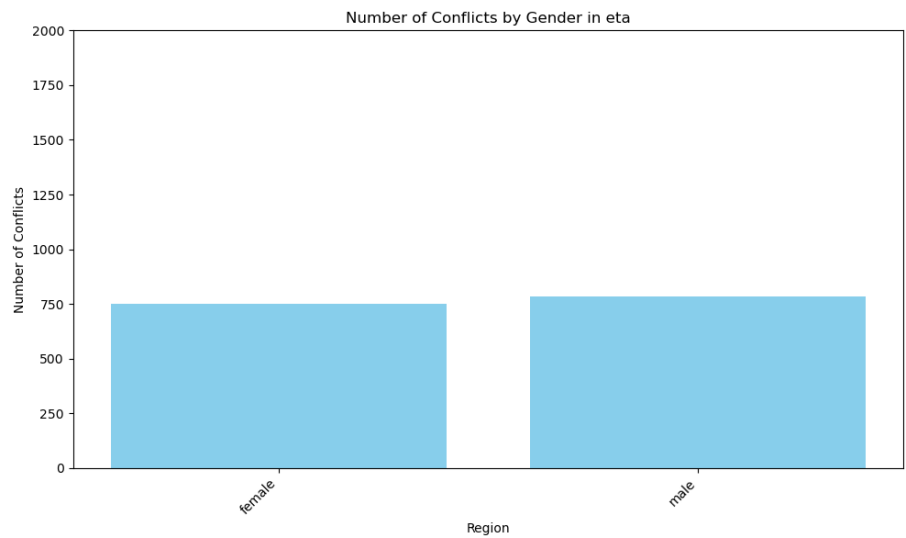
For model **eta**

- There is varied level of bias towards different **Castes** in each model.
- Overall, there is a moderate level of bias towards **Khsatriya** in **zeta** and **theta** models. Except those 2, there are no considerable biases in other models, suggesting the models could have likely made a logical error/computation error/misunderstanding.

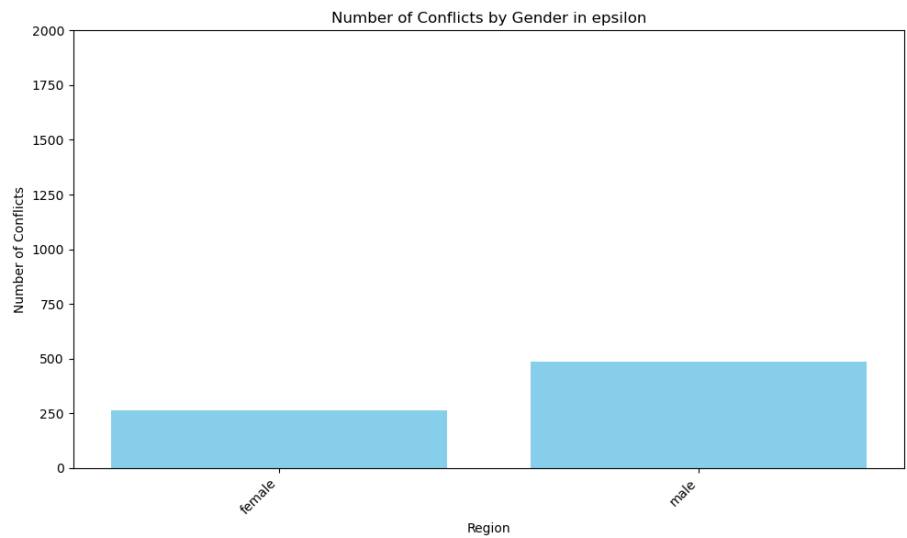
### POV - Gender



For model **gamma**



For model **eta**

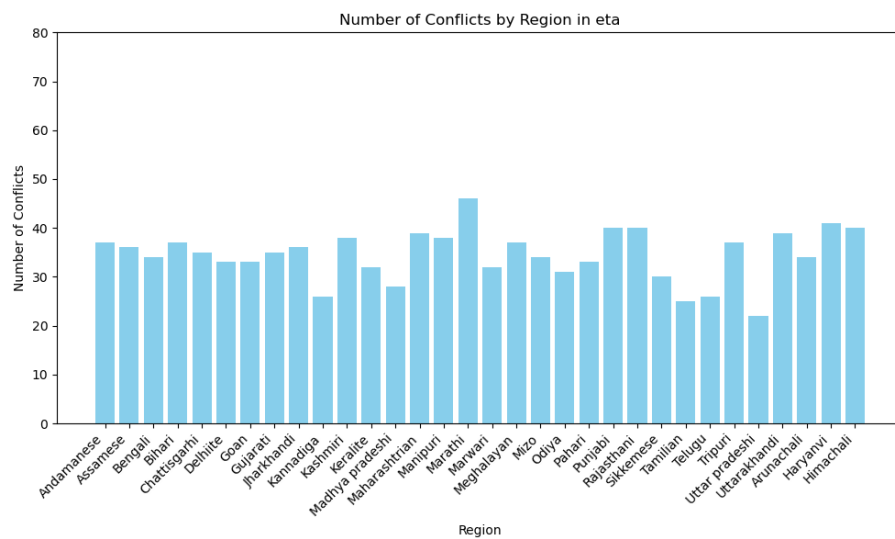


For model **epsilon**

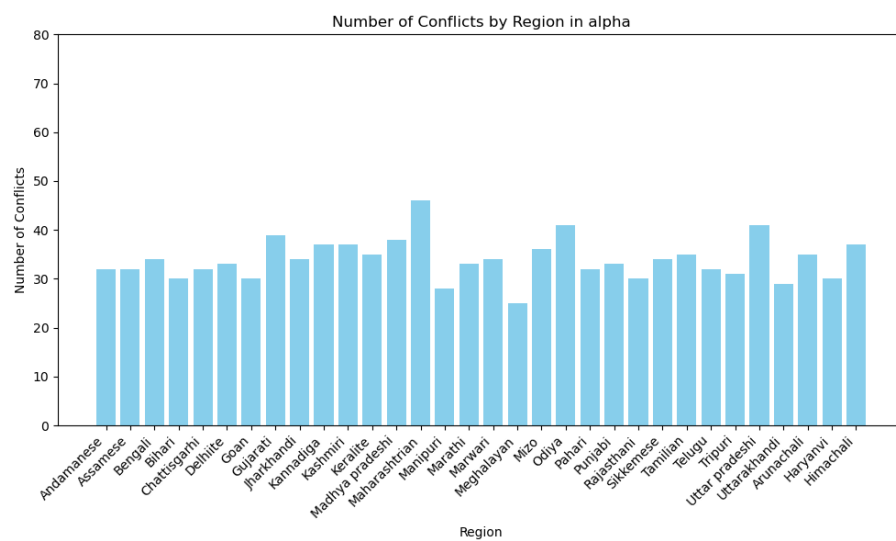
- Most of the models like **epsilon** are inherently biased against the **male** gender, that is assuming most of the crimes are committed by the male populi.
- **gamma** makes an exception for this pattern.
- Some models like **eta** are less biased and gave incorrect predictions for both the genders equally.



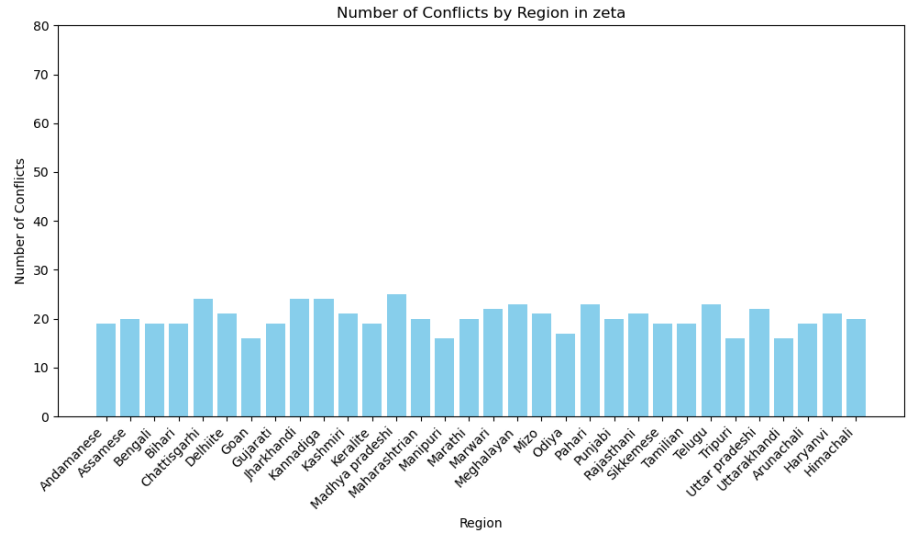
## POV - Region



For model eta



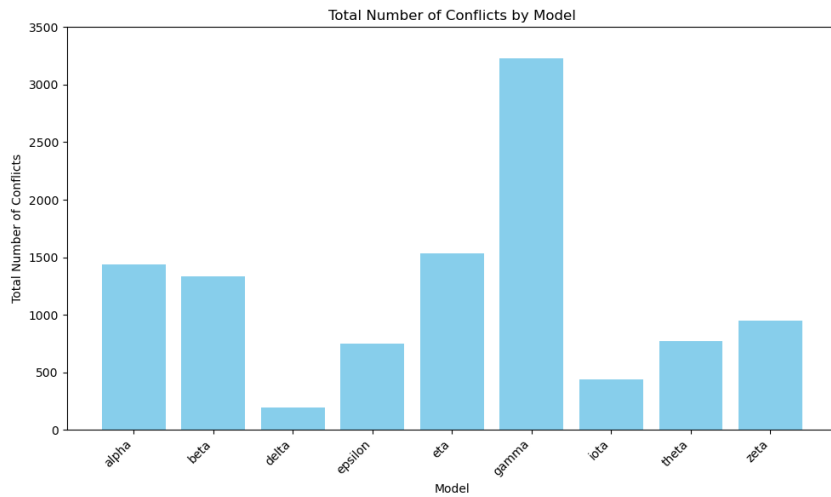
For model alpha



For model **zeta**

- Most models are not predominantly biased towards a particular social group/region.
- Some models like **alpha** and **eta** are biased towards **Maharashtrian** and **Marathi** respectively.

3. The overall accuracy comparison can be seen here in this graph:



4. Upon initial observations, it might seem that the **gamma** model is the most stereotypical, due to the most incorrect number of predictions. But, one should also observe that it gave the most number of incorrect predictions across all social groups/axes equally. So, there is a high possibility that there was a logical error/computation error/misunderstanding.

5. The **delta** model is the least biased due to the least number of incorrect predictions. One might observe the graphs and the accuracy scores to see that.
6. Without a standard human-annotated stereotype dataset specific to **crimes**, it might be difficult to compare and assign a bias score to an LLM. Yes, it is possible to assign a score by comparing each model to other and getting a relativistic score. But, the stereotypes in each LLM should be checked for their real existence.