Annotating Indian Regional Biases using Large Language Models: Evaluation and Analysis

Debasmita Panda[1], Akash Anil[1], Neelesh Kumar Shukla[2]
[1]Department of Data Science and Engineering, Indian Institute of Science Education and Research, Bhopal
[2]Oracle Industries AI, Oracle Corporation
{debasmitap21, anila}@iiserb.ac.in, neelesh.kumar.shukla@oracle.com

Abstract

Social biases based on regional identity (or regional bias) are often observed in Indian contexts on major online social networks and require critical attention. However, due to large linguistic and cultural diversity, high annotation costs, and inherent human biases, very little annotated data exists on regional biases in the Indian context. Recently, Large Language Models (LLMs) have garnered attention for the automatic annotation of text. However, such annotation efforts are largely limited to English texts, and LLMs often perform poorly when applied to low-resource languages. Therefore, this paper focuses on understanding the capabilities and challenges of popular open-source LLMs in annotating Indian regional biases. We utilize the recently proposed `IndRegBias` dataset, which consists of Indian regionally biased social media comments in both English and code-mixed formats. First, we assess the annotation capabilities of LLMs in a zero-shot setting and critically analyze their performance across different writing styles, including code-mixing, transliteration, and English. We find that the majority of LLMs exhibit low agreement with human annotations (measured using Cohen's kappa, κ). Consequently, we extend our study by fine-tuning the models using 50% of the data and evaluating them on the remaining 50%. We observe a significant improvement in annotation agreement (κ) for all the LLMs. To further assess the capabilities of the fine-tuned models, we evaluate them on 500 newly collected social media comments discussing regional issues in India. The results show that most fine-tuned LLMs outperform their zero-shot counterparts when annotating these new comments.

1 Introduction

Regional bias (RB) is a type of social bias that mainly originates from stereotypes toward a particular geographical identity (Tomar et al., 2025; Faisal and Anastasopoulos, 2023; Bhatt et al., 2022). RBs are observed across many countries but are predominant in countries with diverse cultures, linguistic diversity, large populations, and varied geographical regions, such as India (Dev et al., 2023; Khandelwal et al., 2024). Although social biases based on race, gender, and economy have been extensively studied in Natural Language Processing (NLP) research, to the best of our knowledge, very little research has considered regional biases exclusively (Khandelwal et al., 2024; Tomar et al., 2025).

Recently, (Panda et al., 2026) proposed `IndRegBias`, a novel dataset capturing 25,000 social media comments from YouTube[1] and Reddit[2] discussing regional issues in India. `IndRegBias` extracts code-mixed, English, and transliterated comments and employs human annotators to label comments as regional bias or non-regional bias (NRB). Although `IndRegBias` provides a balanced set of RB and NRB examples with the capability to study RBs and their effects on LLMs, it is limited in incorporating Indian regional diversity. This is because (i) human annotators are costly, (ii) achieving high agreement is difficult due to inherent bias or hidden stereotypes, and (iii) code-mixing and transliteration of comments make annotation harder. Thus, to study regional biases at large scale and diversity, there is a requirement to exploit automatic annotators, which can contribute a large volume of examples to generalize NLP models.

To automatically annotate text data, many recent studies (Ding et al., 2023; Törnberg, 2023; Ul Haq et al., 2025) exploit LLMs. These studies show that LLMs are considerably effective as augmented annotators (Alizadeh et al., 2025) and are helpful in generating large synthetic datasets useful for training NLP models (Anikina et al., 2025). However, it is noted in (Jadhav et al., 2025) that LLMs per-

[1]https://www.youtube.com/
[2]https://www.reddit.com/

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 255–263

form better in annotating English texts and struggle with low-resource and code-mixed texts. Further, it is evident from IndRegBias that a significant number of RBs are written in a code-mixed manner, including low-resource Indian languages and transliteration. Thus, it is important to evaluate whether LLMs can be exploited to annotate RBs.

This paper focuses on an extensive evaluation of nine open-source LLMs for automatically annotating comments in IndRegBias. We use zero-shot prompting and compare the annotation agreement of LLMs with the labels in IndRegBias using Cohen's Kappa (κ) (Cohen, 1960). We observe that almost all LLMs show low κ values (below 0.49), indicating that direct inference using LLMs may lead to poor annotation quality.

The majority of recent studies have demonstrated the benefits of fine-tuning LLMs when dealing with low-resource or rare text data (Jadhav et al., 2025). As observed in our zero-shot experiments, we performed low-rank domain adaptation and fine-tuned the LLMs using 50% of the comments (balanced RBs and NRBs) in IndRegBias. We tested the fine-tuned LLMs on the remaining 50% of comments and observed a substantial gain in annotation agreement (κ). This observation indicates that focused domain adaptation may be helpful in creating such challenging datasets with the assistance of LLMs.

To further test this hypothesis, we selected the fine-tuned LLMs and then tested on newly collected YouTube comments following the procedures described in (Panda et al., 2026). We observed consistent improvements in annotation performance for the fine-tuned LLMs compared to their zero-shot counterparts and thus infer that focused fine-tuning is helpful for creating large corpora on RBs.

The major contributions of this paper are:

1. Comprehensive comparison of zero-shot prompting for annotation task over Indian regional biases.

2. Critical analysis on annotation performance and text writing style, i.e., code-mixed, transliterated, and English.

3. Fine-tuning and empirical evaluation on enhancing the corpus using automatic annotations for regional biased comments.

Rest of the paper is organized as follows. Section 2 presents brief descriptions for the related studies, followed by Section 3 which describes the experimental setups and methodologies adopted for this study. Section 4 discusses the empirical investigations along with results and analysis. Finally, Section 5 concludes the paper.

2 Related Studies

Due to the surge in uses of online platforms such as social networks and blogs, there is an enormous amount of data generated (Tomar et al., 2025). Consequently, annotating a large volume of data puts larger cost and time. Therefore, recently NLP research has undergone a paradigm shift through the uses of LLMs as autonomous agents for data annotation (Gilardi et al., 2023). The current trends surrounding the LLMs for annotation emerge from their ability to provide a scalable, cost-effective alternative to human annotations. (Ding et al., 2023) studied benchmarks through performing annotation tasks using ChatGPT-3 and comparing to human annotation, which reports that the GPT model can perform significantly for simple tasks like sentiment analysis. Furthermore, (Törnberg, 2023) evaluated ChatGPT-4 for complex tasks, namely, annotating political affiliations or author's alignment, and reports an impressive performance often outperforming the domain experts. Consequently, there is a surge in the development of sophisticated pipelines like the Multi-LLM Consensus and Human Review (MCHR) framework (Yuan et al., 2025), which leverages independent model voting for the annotation work along with the Human-in-the-Loop framework to curate high-quality datasets. However, (Alizadeh et al., 2025) emphasizes the impact of supervised fine-tuning for the open-source Large Language Models, which provided comparable performance to the proprietary models like GPT 3.5.

While LLMs are efficiently used in annotating English texts, they show limited ability when subjected to low-resource, multilingual, nuanced, and socially situated contexts (Ul Haq et al., 2025; Pavlović and Poesio, 2024). Furthermore, a majority of the LLMs exhibit a neutrality-bias or safety-refusal behavior due to rigid safety guardrails or lack of pretraining with cultural context and might incorrectly label (Jadhav et al., 2025). (Jadhav et al., 2025) reports that for low-resource languages like Marathi, LLMs consistently lag behind the specialized fine-tuned models like MahaBERT. This failure is associated with the lack of high-quality training data and the inherent difficulty models face while

Category	RB	NRB	Total
Code-Mixed	3,405	2,910	6,315
Transliterated	108	355	463
English	9,502	8,720	18,222
Total	**13,015**	**11,985**	**25,000**

Table 1: Distribution of the IndRegBias dataset across Regional Bias (RB) and Non-Regional Bias (NRB) labels categorized into three types of text writing styles, i.e., (i) Code-Mixed: more than one language is used in a comment, (ii) Transliterated: Indian regional languages are transliterated in English, and (iii) English only.

processing non-English syntax. In low-resource NLP tasks, the quality of LLM-generated labels often falls short of human-generated labels, particularly in capturing cultural nuances that do not exist through the generic pre-training predominantly in English-centric training corpora (Nasution and Onan, 2024).

While low-resource languages and social nuances limit the annotation capabilities of LLMs, social media comments capturing regional biases pose more challenges (Bhatt et al., 2022) because of code-mixing and transliteration by many users. Therefore, this paper attempts to study the annotation capabilities of LLMs for the comments in `IndRegBias` and presents a comprehensive analysis with possible solutions using fine-tuning.

3 Experimental Setups and Methodology

In this section, we discuss the dataset, methodology, and experimental setups used to assess the annotation capabilities of LLMs.

3.1 Dataset

We utilize the `IndRegBias` dataset (Panda et al., 2026), a specialized corpus consisting of 25,000 comments expressing 13,015 RBs and 11985 NRBs curated from popular social media platforms, namely Reddit and YouTube. This dataset was specifically designed to capture regional biases (and non-regional biases) in the Indian context. Furthermore, the comments in `IndRegBias` capture code-mixed texts (regional languages such as Hindi, Bangla, etc. along with English) and transliterated text. Therefore, we further categorize the data into three formats: English, Code-Mixed (e.g., Hinglish or Hindi + English, etc.), and Transliterated (Indian languages written in Roman script). As detailed in Table 1, the distribution reveals a predominant English subset, followed by a significant portion of code-mixed texts, which presents a unique challenge for automated annotation task. For the code-mixed category Hinglish (Hindi + English) represents the vast majority as shown in Fig 1. This is expected given its prevalence in digital communication and media in India. Apart from Hindi, Kannada and Bengali show the highest levels of code-mixing with English. There is a small but notable presence of trilingual mixing (e.g., Bengali+Hindi+English), which adds a layer of complexity to NLP tasks like language identification or sentiment analysis. Similar to the code-mixed data (where Hinglish was the most frequent), Hindi is also the primary language used in a transliterated category, along with notable presence of Bengali and Telugu (Refer Figure 1).

3.2 Inter-Annotator Agreement

Similar to `IndRegBias` (Panda et al., 2026), we exploit Cohen's Kappa (κ) coefficient (Cohen, 1960) as the metric to measure the agreement ratio of various LLMs as annotator to the true labels in `IndRegBias`. Cohen's Kappa provides a robust measure of reliability by explicitly accounting for the possibility of agreement occurring by chance. The coefficient is calculated using the following equation:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_o represents the relative observed agreement among raters, and p_e represents the hypothetical probability of chance agreement.

3.3 LLM Evaluation and Analysis Strategies

We used nine open-source LLMs listed in Table 2 for the evaluation of annotating RBs/NRBs. These LLMs were chosen based on their open-source availability and ability to fit in the GPU memory (around 141GB, H200). To evaluate the annotation capability with the true labels in `IndRegBias`, we follow the following strategies:

- **Zero-Shot LLM Annotation on `IndRegBias` (Kojima et al., 2022):** We exploit a suitable LLM prompting technique (refer Box 3.3.1) based on chain-of-thought (Wei et al., 2022) reasoning. In this setting, LLMs annotate the comments as either RB or NRB. These labels are then used to find the agreement with the true labels in `IndRegBias` using Cohen's Kappa (κ).

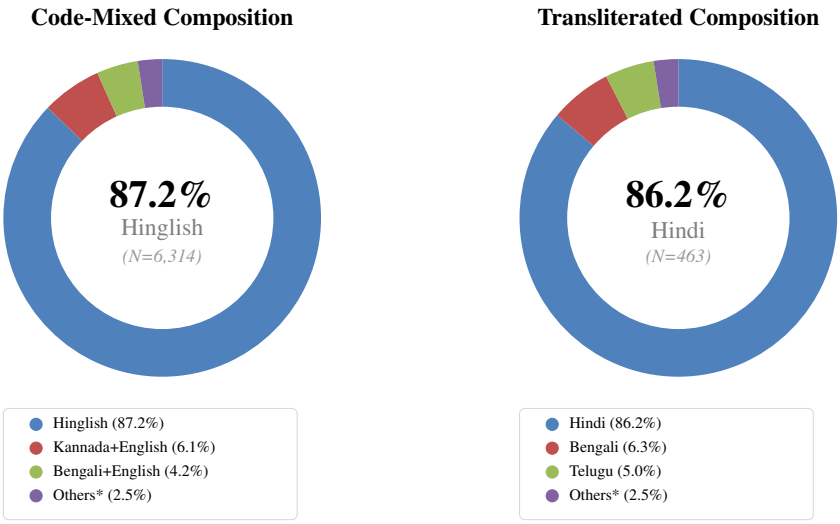

Figure 1: Distribution of Code-Mixed and Transliterated categories within the IndRegBias dataset. Sub-legends provide a structured regional breakdown, including code-mixed (Language+English) along with other combinations as discussed in 3.1 and transliterated comments.

- **Fine-Tuned LLM Annotation on `IndRegBias`:** This setting utilizes supervised fine-tuning approach using the prompt as mentioned in Box 3.3.2. along with the hyper-parameters presented in Table 3. Furthermore, it exploits Parameter Efficient Fine-Tuning (PEFT) (Xu et al., 2026) using the Low-Rank Adaptation (LoRA) (Hu et al., 2021) technique. In each fold, the 25,000 comments are split into 40% for training (10,000), 10% for validation (2,500), and 50% for testing (12,500), maintaining the class balance for the number of RBs and NRBs across all partitions. The five folds are non-overlapping test sets that rotate in such a manner that each comment is evaluated exactly once. We report the mean κ over the five test splits. Additionally, for each fold, we compute zero-shot κ on the same test partition, enabling a controlled comparison between the two strategies on an identical set of data.

- **Fine-Tuned LLM Annotation on Newly Collected YouTube Comments:** In this setting, we collect 500 new comments from YouTube following (Panda et al., 2026). We curated a balanced set of RBs and NRBs and manually annotated these using neutral annotators, qualifying the protocols mentioned in (Panda et al., 2026). We exploit the best fold of fine-tuned LLMs as a pre-trained model and annotate these 500 new comments.

Full Model Name	Shortened Name
Qwen_3_8b	Qwen-8B
Qwen_3_32b	Qwen-32B
DeepSeek-R1-Distill-Qwen-14B	DSeek-14B
Krutrim-2-instruct	Krutrim-2
Mixtral-8x7B-Instruct-v0.1	Mixtral-MoE
LLaMa_3.1_8b	Llama-8B
Mistral_Nemo_12b	MNemo-12B
Mistral_7b_v0.3	Mistral-7B
Gemma_3_27b_it	Gemma-27B

Table 2: Open Source Large Language Models used in this study and their Shortened name to be used in the paper.

Similar to zero-shot and above fine-tuning setups, κ measures the annotation agreement between fine-tuned LLMs and our annotation.

We note that the zero-shot prompt as showcased in Box 3.1 employs a detailed definition of regional bias—mentioning explicitly the hate speech and generalizations based upon culture. This is done to guide models that have no prior exposure to the task, and LLMs rely entirely on the prompt to conduct the annotation task. In contrast, the fine-tuning prompt as presented in Box 3.2, has a broader definition (spanning linguistic, cultural, economic, political, and infrastructural aspects) because the model learns from the labeled examples during training. The underlying annotation task—whether a comment exhibits regional bias—remains identical across both settings; only the degree of definition differs to suit each experiment.

3.1 Zero-Shot Annotation Prompt

Role: You are an expert socio-linguist specializing in Indian regional biases. Your task is to analyze comments to detect if they contain 'Regional Bias' (Label 1) or are 'Non-Biased' (Label 0).

DEFINITIONS
Label 1 (Regional Bias):
The comment contains ANY of the following regarding an Indian state or community: Negative/Positive Stereotypes, Hate Speech/slurs, or Cultural Generalizations.

Label 0 (Non-Regional Biased):
Factual statements, non-related comments, or personal experiences without generalizing.

OUTPUT FORMAT
1. First, engage in a 'Thinking' process to analyze the sentiment.
2. Then, output the final annotation strictly in this format:
`Final Label: [0 or 1]`

`Comment: [INSERT COMMENT HERE]`
`Thinking:`

3.2 Fine-Tuning Input Template

Persona: You are an expert socio-linguist with extensive knowledge of the Indian subcontinent's diverse states, cultures, languages, and regional dynamics. Your task is to annotate comments based on the presence of Regional Bias.

Definitions:
- **REGIONAL BIAS (1):** Comments that reinforce or propagate biases about Indian states, regions, or their people based on linguistic, cultural, economic, political, or infrastructural aspects.
- **NON-REGIONAL BIAS (0):** Comments that are neutral or factual without generalisations, or unrelated to regional characteristics.

`Annotate the following comment as Regional Bias (1) or Non-Regional Bias (0).`

`Comment: [INSERT COMMENT HERE]`
`Annotation: [LABEL]`

4 Results and Analysis

This section discusses the annotation agreement κ shown by various evaluation strategies discussed in Section 3. Furthermore, a comprehensive analysis based on the text writing style (Code-mixing / Transliterated / English) have been provided to understand the empirical investigation at the fine-grained level.

4.1 Zero-Shot Annotation Agreement

Table 4 presents the annotation agreement measured using κ for all the LLMs mentioned in Table 2 with respect to the true labels in `IndRegBias`.

Configuration	Details
Method	LoRA (16-bit BFloat16)
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.05
Target Modules	All Linear Layers (q, k, v, o, gate, up, down)
Epochs	10
Learning Rate	2e-4
LR Scheduler	Cosine (Warmup Ratio: 0.03)
Optimizer	AdamW (8-bit)
Batch Size	8 (dev) $\times$ 4 (acc) = 32 (eff)
Max Length	2048 tokens
Early Stopping	Patience: 3, Threshold: 0.01
Validation	5-Fold Stratified CV
Split Strategy	Rotation Logic
Data Split	40% Train, 10% Val, 50% Test

Table 3: Hyper-parameters for Instruction-Based Supervised Fine-Tuning

Models	Agreement (κ)
Qwen-8B	**0.480**
DSeek-14B	0.478
Krutrim-2	0.460
Mixtral-MoE	0.451
Llama-8B	0.439
Qwen-32B	0.433
MNemo-12B	0.425
Mistral-7B	0.394
Gemma-27B	0.319

Table 4: LLM annotation agreement performance using Cohen's Kappa κ using zero-shot setting on `IndRegBias`

It is evident from Table 4 that all the LLMs achieve poor agreement (κ below 0.49). Furthermore, Qwen-8B performs best out of all the LLMs. Moreover, it should be noted that on the zero-shot annotation tasks, LLMs with larger parameters such as Qwen-32B, MNemo-12B, and Gemma-27B achieve poor agreement when compared to smaller LLMs, i.e., Qwen-8B. On the other hand, while Krutrim-2 is fine-tuned over Indian corpus (referred as Indic Language Model or ILM), it could not achieve a better agreement than LLMs such as Qwen-8B and DSeek-14B. It can be inferred from these observations that for automatic annotation of Indian regional biases, fine-tuning over dataset exclusively discusses regional biases is an important requirement.

Table 5 presents the annotation agreements by the LLMs when the comments in `IndRegBias` were categorized with respect to text-writing styles, i.e., Code-mixed, Transliterated, and English. As per the expectation, it is evident from Table 5 that

Models	Code-Mixed (κ)	Translit. (κ)	English (κ)
Qwen-8B	0.370	**0.387**	0.517
DSeek-14B	0.343	0.385	**0.521**
Krutrim-2	0.367	0.339	0.491
Mixtral-MoE	**0.374**	0.386	0.476
Llama-8B	0.280	0.332	0.491
Qwen-32B	0.329	0.364	0.468
MNemo-12B	0.288	0.327	0.471
Mistral-7B	0.283	0.313	0.431
Gemma-27B	0.287	0.346	0.325

Table 5: Zero-shot performance across different text-writing categories using Cohen's Kappa (κ). Values represent model agreement on the annotation for regional bias / non-regional bias. [Translit.: Transliterated]

Model	ZS (κ)	FT (κ)	% improvement
Qwen-8B	**0.482**	0.791	64.1
DSeek-14B	0.479	0.779	63.0
Krutrim-2	0.458	0.788	72.1
Mixtral-MoE	0.467	**0.798**	70.8
Llama-8B	0.438	0.774	76.7
Qwen-32B	0.434	0.785	80.8
MNemo-12B	0.421	0.783	85.9
Mistral-7B	0.400	0.776	94.0
Gemma-27B	0.322	0.791	>100

Table 6: Comparison of Zero-Shot (ZS) and Fine-Tuning (FT) annotation agreement using Cohen's Kappa (k). 50% of `IndRegBias` has been used for training and validation while remaining 50% has been used for testing the annotation agreement.

all the LLMs (except Gemma-27B) achieve the best annotation agreement for comments written in English. Furthermore, except Krutrim-2, all the LLMs have second best agreement for the comments written using Transliteration. This observation is consistent as Krutrim is finetuned over Indian corpus and shows a better performance in Code-mixed when compared to Transliteration. For category-wise, DSeek-14B performs best in English, Qwen-8B performs best in Transliterated, and Mixtral-MOE performs best in Code-mixed. Overall, Qwen-8B performs competitive in all the categories. These findings underscore the inherent difficulty LLMs face when processing Indian regional biases expressed through Code-mixed and Transliterated texts, highlighting the need for specialized training for automatic annotation.

The better annotation agreement for English comments in Table 5 aligns with their availability i.e., 72.9% (Refer Table 1). However, we note that this observation is not consistent for code-mixed and transliterated comments where we observe that LLMs perform better for Transliterated comments in comparison to Code-mixed even though transliterated comments are much lesser than code-mixed comments. This shows the complexity related to code-mixing writing style and the limitations of LLMs in processing such multilingual texts.

4.2 Fine-Tuned Annotation Agreement

Table 6 presents the average of five-fold stratified annotation agreements by the LLMs using fine-tuning setup discussed in Section 3. Furthermore, Table 6 also presents the average of zero-shot annotation agreement on the same set of test data as used in the corresponding fine-tuning experiment. It is evident from the results that fine-tuning setup consistently improves the annotation agreement for all the LLMs. For instance, the best model in zero-shot setting namely, Qwen-8B gets improved by approximately 64.8% whereas the poorest model in zero-shot, ie., Gemma-27B is improved by more than 100% on annotation agreement. Furthermore, in fine-tuning setup, we observe that Mixtral-MOE gives best annotation agreement, while Qwen-8B and Gemma (surprisingly) perform comparably.

Table 7 provides a detailed comparison between zero-shot and fine-tuning performance across different text-writing categories discussed above (refer Section 3). These results reveal a profound enhancement in the model's reliability after supervised fine-tuning. While results corresponding to zero-shot setting for all the categories remains lower, all the LLMs show significant improvement in annotation agreement after fine-tuning. For example, Mixtral-MoE, achieving the highest agreement scores in both code-mixed (0.6937) and English (0.8327) categories. Qwen3_8b showed the highest performance with kappa value as 0.7944 for Transliterated category. These findings suggest that fine-tuning the models on a domain-specific dataset may lead to enhancement in performance, which becomes essential for navigating complex linguistic nuances, especially for the code-mixed and transliterated comments.

4.3 Fine-Tuned Annotation Agreement on newly collected YouTube Comments

As observed in Tables 6 and 7, the fine-tuning strategy significantly enhances the annotation agreement, motivated with this, we attempted to test the fine-tuned LLMs in Table 6 on the newly collected 500 comments extracted from YouTube. Figure 2 presents a comparative result of this annotation task by fine-tuned LLMs and their corresponding

Models	Code-mixed		Transliterated		English	
	ZS (κ)	FT (κ)	ZS (κ)	FT (κ)	ZS (κ)	FT (κ)
Qwen-8B	0.343	0.674	**0.451**	**0.794**	**0.526**	0.831
Qwen-32B	0.301	0.670	0.386	0.743	0.475	0.825
DSeek-14B	0.313	0.653	0.437	0.725	0.521	0.822
Gemma-27B	0.274	0.677	0.364	0.736	0.328	0.830
Krutrim-2	**0.347**	0.663	0.412	0.766	0.491	0.830
Llama-8B	0.259	0.646	0.400	0.710	0.500	0.819
Mistral-7B	0.259	0.651	0.351	0.713	0.436	0.819
MNemo-12B	0.260	0.663	0.375	0.755	0.469	0.826
Mixtral-MoE	0.343	**0.694**	0.421	0.762	0.486	**0.833**

Table 7: Comparison of Zero-Shot (ZS) and Fine-Tuning (FT) performance across different text-writing categories using Cohen's Kappa κ.

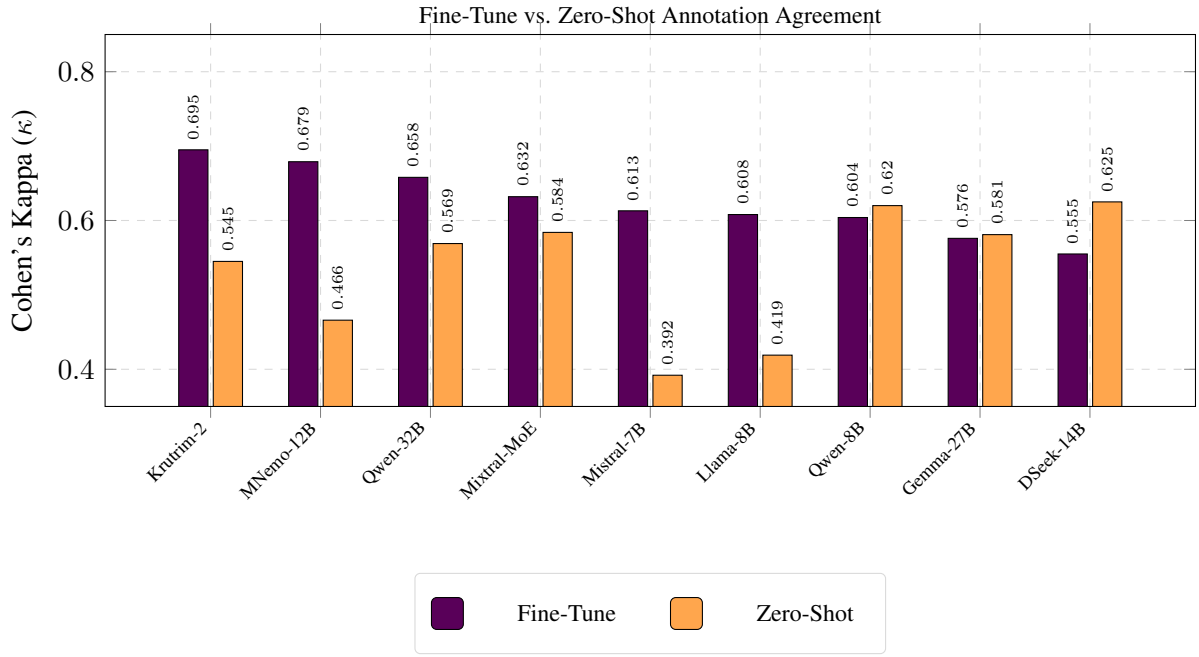

Figure 2: Annotation agreement (κ) for Fine-Tuning and Zero-Shot strategies on the newly collected YouTube comments describing a balanced set for regional and non-regional biases using Cohen's Kappa (κ).

zero-shot counterparts.

From Figure 2, it is evident that while zero-shot inference provides a competitive baseline for some LLMs, supervised fine-tuning is essential for achieving high agreement for annotating the regional bias dataset. To be precise, Krutrim-2 achieved the highest overall performance on the annotation task with a κ score of 0.6950, a significant improvement over its zero-shot counterpart, achieving κ a score of 0.5447. Supervised fine-tuning proved particularly beneficial for models such as MNemo-12B and Llama-8B, which saw substantial gains in agreement scores, rising from moderate levels (0.46 and 0.41, respectively) to the substantial agreement range (above 0.60).

We note that three models namely, Qwen-8B, Gemma-27B, and DSeek-14B, did not outperform their zero-shot results on the newly collected 500 comments from YouTube. This might be due to variance in data distributions and model's limited capability on generalization.

5 Conclusion

This paper focuses on a comprehensive evaluation and analysis of open-source Large Language Models (LLMs) on annotating Indian regional bias dataset, namely, `IndRegBias`. We perform zero-shot and fine-tuning experiments for annotation tasks. From the experiments, it is revealed that almost all the LLMs perform poor on annotation tasks in zero-shot setting. On contrary, fine-tuning LLMs over just 50% of data resulted in a much improved annotation agreement for all the LLMs.

We further evaluated, fine-tuned LLMs for testing a newly collected regionally biased (and approximately equal number of non-regional biased) examples. A majority of the finetuned LLMs perform consistently better than their corresponding zero-shot counterparts.

We critically analyze the performance of annotation by LLMs in three categories of text-writing styles, namely, Code-mixing, Transliterated, and English. It was observed that a majority of the LLMs perform better when comments are written in English whereas they perform poor for code-mixing and transliterated texts.

Our evaluation and analysis reveals that a focused training of LLMs might help in automatic annotation of nuanced and challenging regional biased statements and can be extended in the future[3].

6 Limitations

This paper presents a novel evaluation for LLMs in automatic annotation task for regional biases in Indian context. However, we note some of the important limitations which might have affected the annotation agreement performance by various LLMs. We discuss them below.

- **Imbalanced Representation of Text-Style writing categories**: The `IndRegBias` dataset is heavily skewed toward English (18,222 comments; 72.9%), with Code-mixed texts comprising 25.3% (6,315) and Transliterated texts only 1.9% (463). During Fine-Tuning, the models see fewer Code-Mixed and Transliterated examples in each training fold, limiting their ability to learn the linguistic patterns specific to these styles. This is reflected in Table 7: even after Fine-Tuning, the best Code-Mixed κ (0.694, Mixtral-MoE) remains substantially below the best English κ (0.833, Mixtral-MoE), a gap of nearly 14 points. There is roughly 40:1 ratio between English and Transliterated comments which has led to this consequence. Future work could address this through targeted augmentation of Code-Mixed and Transliterated examples or through stratified oversampling during fine-tuning.

- **Pre-training and Cultural Context Gaps**: Through the categorization process, we have found a variety of combinations of Indian languages with English texts along with the cultural context. This is where the models' pretraining information comes into the picture, as they have been trained in a general setting for any language. They might ignore the code-mixed part because they have not been pretrained on this type of data, resulting in poor performance. On the other hand, if we consider social biases through comments, they shall consist of code-mixed text-style. In the transliterated category, where most of the language was written in English script, the pretraining on global data played a major role. It is because most of the LLMs that we have tested have not been pretrained on such Indian languages, e.g., DeepSeek, Llama, and Mistral, which might have led to poor performance. Furthermore, models that are pretrained on Indian languages, such as Qwen and Krutrim, show poor performance, possibly due to their general training, which lacks the nuanced cultural context understanding.

- **Evaluation Scale for the newly collected YouTube comments**: The newly collected comments from YouTube comprise 500 comments only, which is significantly less than the 25,000 comments in the original dataset. This set followed the annotation policy from (Panda et al., 2026) as used to prepare the dataset `IndRegBias`. We have balanced instances for both the classes of Regional Bias (RB) and Non-Regional Bias (NRB). More importantly the goal of this evaluation is not to pin down the exact performance of the models, we wanted to answer a simpler question: does fine-tuning consistently help over zero-shot inference when the data comes from outside the training distribution? Six out of nine models show clear improvement, and the three that do not can be traced to specific factors discussed in subsection 4.3, which gives us reasonable confidence in the answer. Of course, a larger and more diverse out-of-distribution evaluation spanning additional regional topics and languages would make these conclusions more robust, and we plan to pursue this in future work.

[3]Code is available at: https://github.com/debby123p/Annotating-Indian-Regional-Biases-using-Large-Language-Models-Evaluation-and-Analysis

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2025. Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1):17.

Tatiana Anikina, Jan Cegin, Jakub Simko, and Simon Ostermann. 2025. A rigorous evaluation of LLM data generation strategies for low-resource languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1142–1158. Association for Computational Linguistics.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing 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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023. Building socioculturally inclusive stereotype resources with community engagement. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195. Association for Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2023. Geographic and geopolitical biases of language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 139–163.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2025. On limitations of LLM as annotator for low resource languages. In *Proceedings of the International Conference on Natural Language and Speech Processing (ICNLSP 2025)*.

Khyati Khandelwal, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott A. Hale. 2024. IndianBhED: A dataset for measuring India-centric biases in large language models. In *Proceedings of the 2024 International Conference on Information Technology for Social Good (GoodIT '24)*, pages 231–239. ACM.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Arbi Haza Nasution and Aytuğ Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*, 12:71876–71900.

Debasmita Panda, Akash Anil, and Neelesh Kumar Shukla. 2026. IndRegBias: A dataset for studying indian regional biases in english and code-mixed social media comments. *arXiv preprint arXiv:2601.06477*.

Maja Pavlović and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110. ELRA and ICCL.

Aditya Tomar, Nihar Ranjan Sahoo, and Pushpak Bhattacharyya. 2025. BharatBBQ: A multilingual bias benchmark for question answering in the indian context. *arXiv preprint arXiv:2508.07090*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Muhammad Uzair Ul Haq, Hussain Ahmad Zafar, and David Broneske. 2025. LLMs as data annotators: How close are we to human performance? *arXiv preprint arXiv:2504.15022*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Lingling Xu, Haoran Xie, S. Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2026. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–20.

Mingyue Yuan, Jieshan Chen, Zhenchang Xing, and Gelareh Mohammadi. 2025. A case study of scalable content annotation using multi-LLM consensus and human review. In *Proceedings of the Generative AI and HCI Workshop at CHI 2025*.