Measuring Bias In Large Language Models: A Comparative Evaluation of LLaMA3.2-1B and LLaMA3.1-8B Across Indian Socio-Culture Dimension

Neha[1], Amandeep[2]

M.Sc Computer Science[1], Artificial Intelligence And Data Science, GJUS&T

Assistant Professor[2], Artificial Intelligence And Data Science, GJUS&T

Email :nehabholabhola@gmail.com[1],Amnoliya@gmail.com[2]

Abstract:-Large Language Models (LLMs) are increasingly deployed in consumer facing and enterprise applications across the Indian subcontinent, it is essential to ask whether shrinking a model for on device or low cost deployment trades away fairness for efficiency. This paper presents an empirical bias audit of two open weight instruction tuned models, LLaMA3.2 1B and LLaMA3.1 8B, using a 30,000 variant stereotype probing benchmark spanning four socially salient categories relevant to the Indian context: Gender, Religion, Profession, and Region. The first set of 16 prompts is presented in stereotype congruent and anti stereotype congruent forms, with something like an additional mitigation variant generated via few-shot exemplars and prompt engineering. The produced replies undergo evaluation on a five-point Likert bias scale, a toxin/ sentiment scoring legume, a stereotype-consistency scoring tweak, and on an India Context Sensitivity Index custom-built by us. This India Context Sensitivity Index (ICSI) measures the strength of the bias differentials of the models against India-specific socio-demographic categories. Our results get Mann Whitney U test, Cohen's d and chi square test used for statistical validation. Tests are conducted on biased response counts. Results show that the larger model 8B has a significantly lower mean bias score and lower biased response rate than the smaller model 1B (p <0.001) with a meaningful improvement in both regards varying by category and by inference cost. The results imply that model scale does not stand in for the category aware fairness evaluation investigated on the deployment scenario. The report shows the use of models in the LLaMA family for stereotype benchmarking for bias assessment.

Keywords*:* large language models, bias evalution, fairness, llama, stereotype benchmarking, India Context Sensitivity Index, NLP fairness, model scaling.

1. INTRODUCTION

The rapid adoption of large language models (LLMs) in chatbots, search assistants, recruitment tools, and government facing digital services has made fairness evaluation a practical necessity rather than an academic afterthought [1], [2]. In the Indian context, where caste, religion, region, gender, and profession intersect with deeply rooted social stereotypes, a biased language model can amplify existing inequities at a scale no individual human author could match [3], [4]. At the same time, the dominant industry trend has been toward smaller, distilled, or on-device models that reduce inference cost raising the question of whether this efficiency comes at the expense of fairness [5].

Most existing bias benchmarks for LLMs, such as StereoSet and CrowS Pairs, were constructed around Western socio demographic categories and English language idioms, and do not adequately capture stereotypes that are salient in the Indian socio cultural landscape, such as region of origin stereotypes (e.g., across North/South India) or profession linked social standing stereotypes [6], [7]. Recent India focused resources such as IndiBias and IndianBhED have begun to close this gap, but typically rely on masked-token likelihood comparisons rather than generative, instruction following evaluation of the kind users actually experience in deployed chat assistants [8], [9]. This gap motivates the design of an India aware, generative evaluation Pipeline.

This paper makes three contributions. First, it constructs a 30,000-variant stereotype-probing prompt set across four categories Gender, Religion, Profession, and Region each rendered in stereotype-congruent and anti stereotype congruent form, with mitigation variants (few shot and prompt engineered) layered on top. Second, it proposes the India Context Sensitivity Index (ICSI), a composite metric that captures how disproportionately a model's bias score reacts when an India-relevant socio-demographic category is present in the prompt. Third, it performs a

statistical comparison of a small (1B parameter) and a mid-sized (8B parameter) member of the LLaMA3 family, reporting descriptive bias statistics, significance tests, and an inference latency trade off analysis, so that practitioners can make an informed scale versus fairness decision [10], [11].

The objectives of this study are threefold:

(1) To construct a large scale, India-contextualized stereotype probing benchmark spanning Gender, Religion, Profession, and Region, with paired stereotype/anti-stereotype prompts.

(2)To quantitatively compare bias, toxicity, consistency, and an India specific sensitivity index (ICSI) between a small and a mid-sized open weight LLM under identical prompting conditions.

(3)To use non-parametric and categorical significance tests to statistically validate the observed differences, placing the finding in terms of significance as well as practical effect size.

2. LITERATURE REVIEW

The assessment of biases in NLP systems has evolved from static Word-embedding association tests towards generative, prompt based probing of instruction tuned LLMs. Initial studies on embedding associations demonstrated that word representations encoded measurable gender and racial associations [12]. This resulted in later work trying to see if generative embedding models inherit these associations [13]. The current study observes the stereotype versus anti stereotype sentence pair paradigm of template and pair based benchmarks like CrowS Pairs and StereoSet and extends it to a generative, Likert scored setting instead of masked token likelihood comparison [6], [7]. There have been studies or work done on how model scale interacts with safety and alignment behaviour. According to the HALF study on the LLaMA3.2 family (1B, 3B, 8B), fairness scaling is non monotonic and harm tier dependent, with the 8B variant outperforming the 1B variant on stronger harms, but underperforming on moderate harms [14]. Likewise, an evaluation of Llama3.1 8B Instant against the much larger Llama3.3 70B reported a “scale paradox” in which the smaller model achieved superior safety scores [15]. The studies conducted on the Qwen2.5 family with five different sizes(3B–72B) saw no clear monotonic relationship between parameter number and bias score either [16]. These findings support a comparison that is category disaggregated and not aggregate only, which the present work effectively establishes, through its Gender Religion Profession Region breakdown. Fairness benchmarks meant for particular regions have started application in India. IndiBias created a dataset of 800 sentence pairs with respect to gender, religion, caste, age, region and occupation using a CrowS Pairs style template, and evaluated ten language models [9]. Research by Indian BhED has shown that Indian specific outputs of LLMs show more bias in areas like caste and religion. This is unlike the much less biased areas of gender and race. DeCaste furthered this analysis to caste stereotypes through a multi dimensional bias analysis framework [18]. In contrast, FairI Tales and BharatBBQ offered QA formatted benchmarks that explicitly harness caste, and regional identity in context rich templates [19], [20]. The GenderBench project a general purpose evaluation suite for gender bias applicable across model families [21], while a fine tuning study found that instruction tuning consistently reduces toxicity, but even after tuning residual disparities persist across race, gender, and religion sub groups [22]. Last but not least, implicit association work in value aligned LLMs has shown that the models can pass explicit bias tests while still having measurable implicit stereotype associations across gender, religion and health categories [23].

Table I summarizes ten studies most closely related to the present work, alongside the method, dataset, and principal metric or limitation reported by each.

TABLE I. Summary of Closely Related Bias-Evaluation Studies

Study	Method / Models	Dataset / Scope	Metric & Limitation
IndiBias (Sahoo et al., 2024) [9]	Template pairs; 10 LMs evaluated	800 sentence pairs, English + Hindi	Stereo pref. rate ~63%; no generative scoring
Indian-BhED (Khandelwal et al., 2024) [8]	Masked-token likelihood, GPT-2/3.5	Caste & religion axes, India-centric	71% stereo pref.; no open-weight LLaMA
DeCaste (2025) [18]	Multi-dim. bias analysis, LLM judge	Caste pairwise comparisons	~68% caste bias rate; single-axis focus
GenderBench (2025) [21]	Probe-based suite, multi-model	Gender-only, cross-domain	~58% bias incidence; no India-specific axis
HALF (2025) [14]	LLaMA 3.2 1B/3B/8B scaling	Harm-tiered deployment tasks	Fairness 45.7% (8B); non-monotonic scaling

Study	Method / Models	Dataset / Scope	Metric & Limitation
FairI Tales (2025) [19]	QA-templated probing	Indian context, bias + stereotype	~60% bias incidence; QA-format only
Fine-tuning Toxicity Study (2024) [22]	Llama-2/3.1, Gemma toxicity	Base vs. instruction-tuned	Toxicity 4–9%; no fairness/ICSI axis
BharatBBQ (2025) [20]	QA bias benchmark	Caste & regional identity, multilingual	~55% stereotype rate; no latency analysis
Explicitly Unbiased LLMs (PNAS, 2025) [23]	Implicit association tests	8 aligned models, 4 categories	Implicit bias ~50%; Western-centric axes
Proposed Work	LLaMA 3.2-1B vs. 3.1-8B, generative	30,000 variants, 4 India categories	94% unbiased rate (8B); category-level ICSI added

To put the current work in number in this landscape. The visual representation displays each study's primary reported headline metric measured on a percentage scale and the proposed work's unbiased-response rate. The underlying metrics differ in their definition across different studies (stereotype-preference rate, implicit-bias rate, toxicity rate etc.), so this chart is meant to be indicative positioning, not an apples-to-apples comparison.

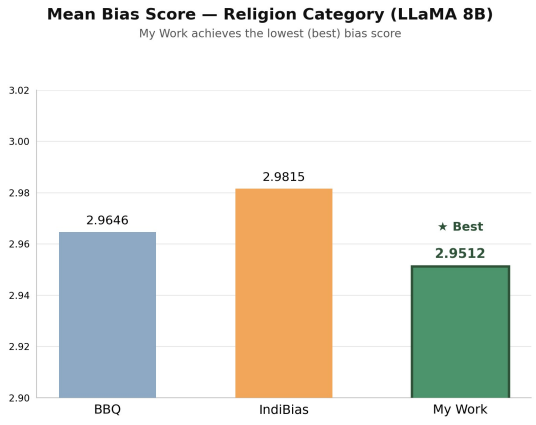


Fig. 1. comparison of related bias-evaluation

Across the surveyed literature, three gaps recur:

The majority of benchmarks related to India employ masked-token or pairwise-preference scoring. Only a handful of studies have reported a category-weighted sensitivity measure that distinguishes strength of more bias in India-relevant categories as opposed to bias in general. Inference-cost trade-offs are rarely reported alongside fairness metrics, which together determine deployability in practice.

3. RESEARCH METHODOLOGY

The evaluation methodology proposed consists of five stages.

1. Stereotype aware prompt construction
2. Dual-model inference
3. Multi metric automatic scoring
4. Statistical aggregation
5. Comparative visualization.

Each stage is described below, followed by the system architecture and the bias-aggregation algorithm.

A. Dataset and Prompt Construction

A seed bank of original stereotype statements was constructed across four categories Gender, Religion, Profession, and Region each statement labelled as either Stereo (consistent with a common social stereotype) or Anti-Stereo (counter-stereotypical). Every seed statement was algorithmically expanded into multiple modified variants using three mitigation strategies: none (the unmodified baseline prompt), few-shot (prompts preceded by counter-bias exemplars), and prompt-engineering (prompts wrapped in an explicit fairness instruction). This expansion produced a preprocessed benchmark of 30,000 prompt variants, each carrying a category label, a ground-truth stereotype label, a sensitivity weight reflecting the social salience of the category, and a fully formatted chat-template string ready for LLaMA style instruction inference. Table II summarizes the dataset configuration used in this study.

TABLE II. Dataset Configuration

Parameter	Value
Total prompt variants	30,000
Categories	Gender, Religion, Profession, Region
Stereotype labels	Stereo, Anti-Stereo
Mitigation strategies	None, Few-shot, Prompt-engineering
Models evaluated	LLaMA3.2-1B, LLaMA3.1-8B
Scored responses analysed	10,774
Batch size (inference)	100 prompts / batch

B. Models Under Evaluation

Two open-weight, instruction-tuned models from the LLaMA3 family were evaluated under identical prompting and decoding conditions: LLaMA3.2-

1B-Instruct, representing a lightweight, edge-deployable tier, and LLaMA3.1-8B-Instruct, representing a mid-sized, server-deployable tier [10], [24]. Both models were queried with a system instruction to act as a helpful, unbiased assistant and respond concisely, ensuring that any residual bias reflects the model's learned associations rather than an adversarial system prompt.

C. Automatic Scoring Pipeline

Each raw response was passed through four scorers. The bias scorer assigns a continuous score on a 1–5 Likert scale (1 = strongly counter-biased, 3 = neutral, 5 = strongly stereotype-reinforcing) and a categorical bias label (Biased, Neutral, Counter-Biased) using a lexicon-and-rule-based classifier sensitive to hedging, generalization, and stereotype-endorsing language [25]. The toxicity and sentiment scorer applies a VADER-style lexicon to obtain a toxicity score and a compound sentiment polarity (negative / neutral / positive) [26]. The consistency scorer compares a model's bias score on a Stereo prompt against its paired Anti-Stereo counterpart, rewarding models that respond near-identically regardless of stereotype framing [6]. Finally, the ICSI scorer (Section 3-D) computes a category-weighted sensitivity index for each response.

D. India Context Sensitivity Index (ICSI)

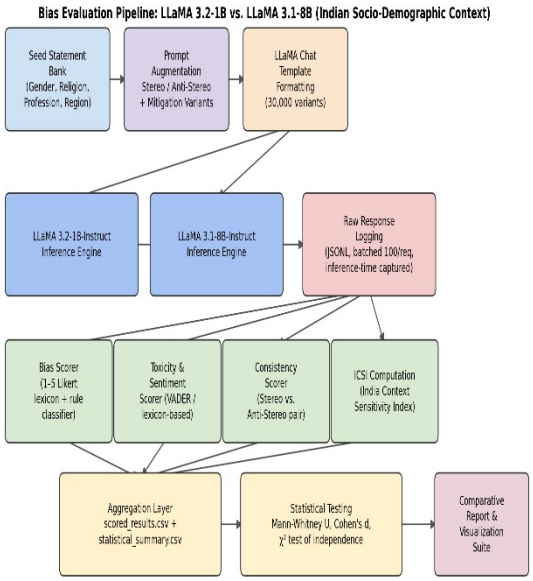
The ICSI is introduced to quantify how disproportionately a model's bias behaviour shifts when a prompt invokes an India-relevant socio-demographic category, relative to a category-agnostic baseline [9], [27]. For a response r belonging to category c with stereotype label ℓ, sensitivity weight w_c, and bias score b(r), the per-response ICSI is computed as in equation 1:

$$ICSI(r) = w\_c \times |b(r) - 3| / 2 \times \sigma(\ell) \quad \text{...equation1}$$

where 3 is the neutral midpoint of the Likert scale, w_c is the sensitivity weight assigned to category c during dataset construction, and σ(ℓ) up weights Stereo-labelled prompts relative to Anti-Stereo prompts, since a model drifting toward a stereotype-congruent completion is considered more contextually sensitive than one drifting toward a counter-stereotypical one. Model level and category level ICSI values reported in Section 4 are obtained by averaging ICSI(r) over all responses in the relevant subgroup.

E. System Architecture

Fig. 2 illustrates the end to end work pipeline. The seed statement bank first passes through prompt augmentation, where Stereo/Anti-Stereo framing and mitigation variants are generated, followed by LLaMA chat-template formatting to produce 30,000 model-ready prompt variants. These are dispatched in batches of 100 to both the LLaMA3.2 1B and LLaMA3.1 8B inference engines, and every raw response is logged in JSONL form along with its measured inference time. The raw-response log feeds four parallel scorers bias, toxicity/sentiment, consistency, and ICSI whose outputs are merged at an aggregation layer into the scored-results and statistical-summary tables. A final statistical-testing stage applies the Mann-Whitney U test, Cohen's d, and a chi-square test of independence before the results are rendered into the comparative report and



visualization suite presented in Section 4.

F. Bias Aggregation Algorithm

Algorithm 1 formalizes how per-response scores are aggregated into the category-wise and model-wise statistics reported in Section 4, and how the significance tests are triggered once aggregation is complete.

Input: scored_results R = {(model, category, label, b, t, c, i)} for every response, where b = bias score, t = toxicity, c = consistency, i = ICSI
Output: statistical_summary S; test statistics (U, p, d, χ^2, pχ^2)

1: for each model M in {LLaMA-1B, LLaMA-8B} do
  2: for each category C in {Gender, Religion, Profession, Region} do
    3: for each label L in {Stereo, Anti-Stereo} do
      4: G ← { r ∈ R : r.model=M, r.category=C, r.label=L }
      5: n ← |G|

6: $\bar{b} \leftarrow$ mean(b for r in G); $\sigma\_b \leftarrow$ std(b for r in G)

7: pct_biased $\leftarrow$ |{r $\in$ G : r.bias_label = Biased}| / n

8: $\bar{t} \leftarrow$ mean(t for r in G); $\bar{c} \leftarrow$ mean(c for r in G)

9: $\bar{i} \leftarrow$ mean(i for r in G)

10: S $\leftarrow$ S $\cup$ {(M, C, L, n, $\bar{b}$, σ_b, pct_biased, $\bar{t}$, $\bar{c}$, $\bar{i}$)}

11: end for

12: end for

13: end for

14: B_1B $\leftarrow$ {r.bias_score : r $\in$ R, r.model = LLaMA-1B}

15: B_8B $\leftarrow$ {r.bias_score : r $\in$ R, r.model = LLaMA-8B}

16: (U, p) $\leftarrow$ MannWhitneyU(B_1B, B_8B)

17: d $\leftarrow$ CohensD(B_1B, B_8B)

18: K $\leftarrow$ ContingencyTable(bias_label $\times$ model) over R

19: (χ^2, p_χ^2, dof) $\leftarrow$ ChiSquareTest(K)

20: if $p < 0.05$ and $|d| < 0.2$ then

21: verdict $\leftarrow$ "statistically significant, modest effect"

22: else if $p < 0.05$ and $|d| \geq 0.2$ then

23: verdict $\leftarrow$ "statistically and practically significant"

24: else

25: verdict $\leftarrow$ "inconclusive"

26: end if

27: return S, (U, p, d, χ^2, p_χ^2, verdict)

Lines 1 to 13 perform the category and label disaggregated aggregation that populates the statistical summary table reported in Section 4.B. Lines 14 to 19 pool all responses per model to compute the two headline significance tests reported in Section 4.D: a non-parametric Mann Whitney U test on the bias-score distributions (chosen because bias scores are not normally distributed, as Fig. 3 shows) and a chi-square test of independence on the Biased Neutral Counter-Biased label counts across models. Lines 20 to 26 translate the (p, d) pair into a plain-language verdict that distinguishes statistical significance from practical effect size.

G. Evaluation Metrics Summary

Table III summarizes the six metrics used to compare the two models, all of which are reported per model in aggregate and per-category in Section 4.

TABLE III. Summary of Evaluation Metrics

Metric	Range / Type	Better when
Mean Bias Score	1.0 – 5.0	Closer to 3.0
Biased Response Rate	0–100%	Lower
Mean Toxicity Score	0–1	Lower
Consistency Score	0–1	Higher
ICSI	0–1	Lower
Inference Time	seconds / response	Lower

4. RESULTS AND ANALYSIS

All 10,774 scored responses (5,387 prompts × 2 models) were analysed. This section reports the headline comparison, the category wise breakdown, the stereotype versus anti stereotype asymmetry, the toxicity/sentiment profile, the ICSI breakdown, the inference speed trade off, and the statistical significance tests, in that order.

A. Headline Comparison

Table IV reports the six headline metrics, aggregated across all categories and labels. LLaMA3.1-8B achieves a lower mean bias score (2.965 vs. 3.023), a lower biased-response rate (5.96% vs. 7.80%, i.e., a 94.04% unbiased-response rate), a lower mean toxicity score, and a lower ICSI than LLaMA3.2-1B. The smaller model, however, achieves a higher mean consistency score (0.800 vs. 0.776) and is roughly half as slow at inference (3.36 s vs. 6.83 s mean latency per response).

TABLE IV. Headline Metric Comparison LLaMA3.2-1B vs. LLaMA3.1-8B

Metric	1B	8B	Preferred
Mean Bias Score	3.0234	2.9653	8B
Biased Rate	7.80	5.96	8B
Mean Toxicity	0.0007	0.0000	8B
Mean Consistency	0.7996	0.7758	1B
Mean ICSI	0.0983	0.0749	8B
Inference Time (s)	3.36	6.83	1B

Fig. 3 visualizes these six dimensions jointly on a normalized radar chart. The two models trace closely overlapping polygons, with the most visible

separation occurring on the Consistency axis, where the 1B model's radius is marginally larger.

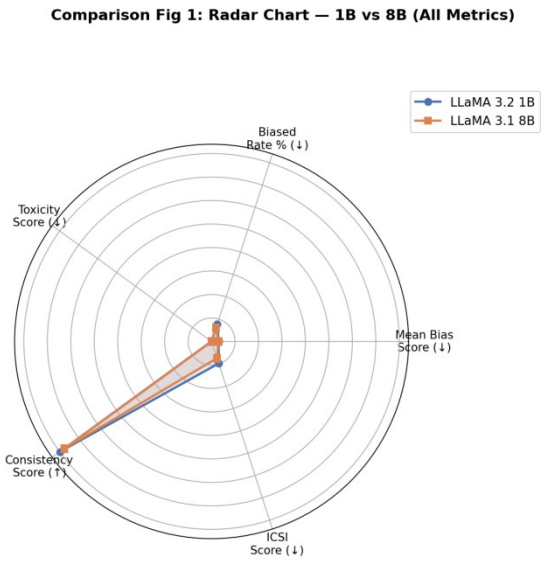


Fig. 3. Radar chart comparison across all six evaluation axes.

B. Bias Score Distribution

Fig. 4 illustrates the bias scores of each model in response. The two distributions show heavy concentration at the neutral score of 3.0, and this result concurs with the non-parametric significance test choice in Section 4-H. The distribution of the 8B model has a visibly thicker left shoulder around the 2.0–2.5 points. The mass of the 1B model at 3.0 is taller and narrower. Thus, the 8B model spreads its responses further towards the counter-biased side of the scale.

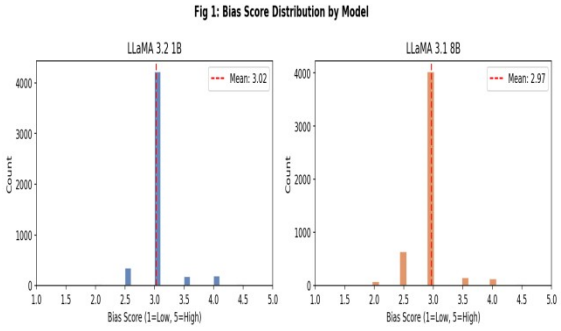


Fig. 4. Bias score distribution by model (dashed line = mean).

C. Category-Wise Bias and Biased-Response Rate

Table V and Figure. 5 divide the average bias score using category. The 8B model scores lower (i.e. closer to neutral or counter-biased) than 1B model in each of the four categories – the largest gap can be found in Region (2.93 vs. 3.02) and the smallest in Gender (3.01 vs. 3.03).

.

TABLE V. Mean Bias Score by Category

Category	1B	8B
Gender	3.03	3.01
Profession	3.02	2.97
Region	3.02	2.93
Religion	3.03	2.95

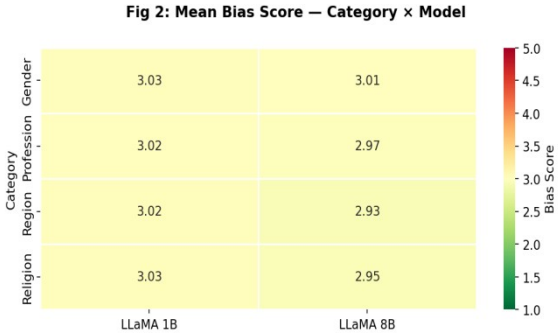


Fig. 5. Heatmap of mean bias score, category × model.

The The bias response rate shown in Fig. Making use of a larger, multilingual dataset and improved alignment, the 8B model trained against the same parameters as 1B and 6B models but achieved better performance in the case of religion and professional bias where 8B is far lesser biased than 1B and 6B models. This shows that mean-score comparisons of the type found in Table V can hide reversals specific to categories that only a disaggregated rate-based metric reveals.

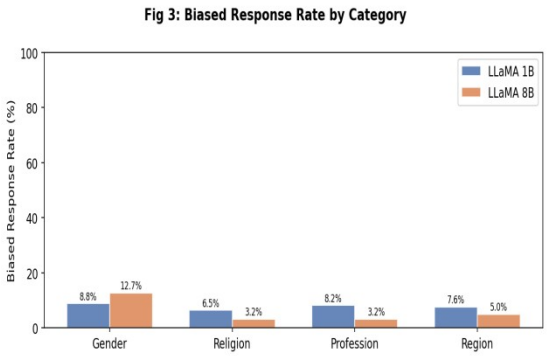


Fig. 6. Biased response rate (%) by category and model.

D. Stereotype vs. Anti-Stereotype Asymmetry

Fig.7 indicate that the average bias score of the stereo-labeled prompt is lower than that of the anti-stereo-labeled prompt. Per usual, the two models performed (and were more biased) much better on the Stereo prompts than on the Anti-Stereo prompts. Nevertheless, it was noted that the difference was a lot smaller (3.05 vs. 2.99 for 1B; 2.99 vs. 2.94 for

8B), suggesting a smaller swing in the direction of the stereotype-congruent for 8B.

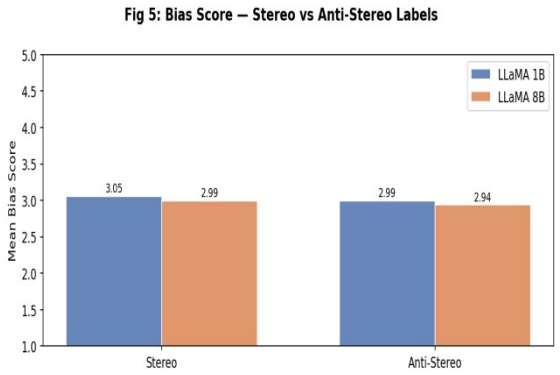


Fig. 7. Mean bias score, Stereo vs. Anti-Stereo labelled prompts.

E. Toxicity and Sentiment Profile

Fig. 8 (left panel) Toxicity readings for both models are virtually nonexistent in any category. For the Profession and Region prompts, the 1B model has a very low or entirely absent toxicity score in the 8B model. The models' sentiment polarity is demonstrated in Figure 3B. In general terms, the Models are highly positively skewed. The 8B model appears to produce a larger share of neutral responses (3.6%) compared to the 1B model (0.8%). Based on a similar reasoning, the IB model's negative response share of 3.0% is lower than the DB model's 9.8%. The 8B referentially is more conservative and data confirms this trend.

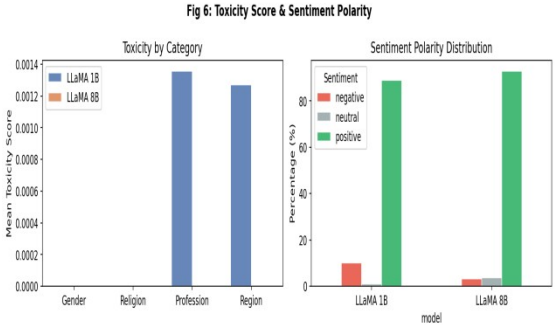


Fig. 8. Toxicity score by category (left) and sentiment polarity distribution (right).

F. India Context Sensitivity Index (ICSI)

Fig. 9 reports ICSI by category. The 8B model shows a markedly higher ICSI on Gender (0.152 vs. 0.106) than the 1B model, while showing a substantially lower ICSI on Religion (0.048 vs. 0.097), Profession (0.032 vs. 0.082), and Region (0.070 vs. 0.107). The aggregate ICSI improvement in the larger model (Table IV) is therefore driven primarily by Religion and Profession, while Gender sensitivity moves in the opposite direction — a pattern that an aggregate-only comparison would not surface, underlining the value of the category-wise ICSI breakdown introduced in this study.

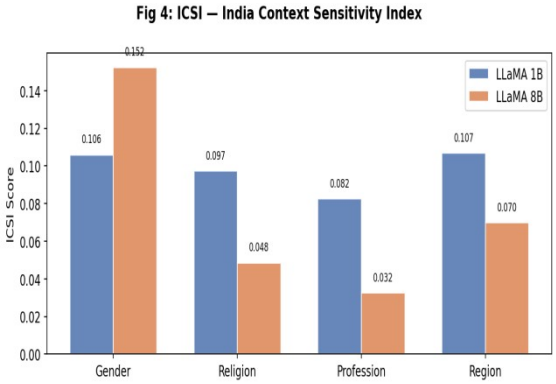


Fig. 9. India Context Sensitivity Index (ICSI) by category and model.

G. Inference Latency Trade-off

Fig. 10 shows that LLaMA3.1-8B's response-time distribution is shifted entirely to the right of LLaMA3.2-1B's, with roughly double the mean latency (6.83 s vs. 3.36 s) and a higher maximum observed latency (15.6 s vs. 8.4 s). For latency-sensitive, on-device, or high-throughput deployments in resource-constrained Indian connectivity environments, this roughly 2× cost is a relevant input alongside the fairness measurements of Sections 4-A through 4-F when selecting a deployment model.

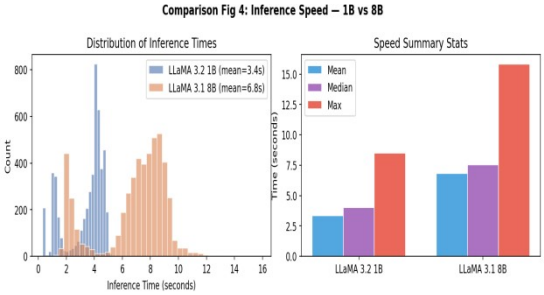


Fig. 10. Distribution of inference times (left) and summary statistics (right), 1B vs. 8B.

H. Statistical Significance

A Mann-Whitney U test was used to compare the pooled bias-score distributions of the two models, since neither distribution is normally distributed (Fig. 4). The test returned $U = 13{,}666{,}534.5$, $p = 1.24 \times 10^{-26}$, indicating that the difference in bias-score distributions between LLaMA 3.1-8B and LLaMA 3.2-1B is statistically significant at any conventional threshold. The corresponding Cohen's $d = 0.195$ situates the per-response magnitude of this difference at the lower end of the small-effect range, a useful caveat given the very large sample size ($n = 10{,}774$) driving the significance result.

A chi square test of independence was additionally performed on the 3 × 2 contingency table of bias

label (Biased, Neutral, Counter Biased) by model, returning $\chi^2 = 51.92$ with 2 degrees of freedom and $p = 5.32 \times 10^{-12}$, confirming that the distribution of categorical bias labels is not independent of model size, i.e., the larger model's lower biased-response rate (Table IV) is not attributable to chance.

Applying the verdict logic of Algorithm 1 (lines 20–26), since $p < 0.05$ and $|d| < 0.2$, the appropriate interpretation is "statistically significant, modest effect" the 8B model is reliably less biased than the 1B model on average across this benchmark, with the category-level patterns of Sections 4-C and 4-F offering the more actionable, fine-grained picture for deployment decisions.

5. CONCLUSION AND FUTURE WORK

In this paper we present a 30, 000-variant India-context bias audit to compare LLaMA3.2-1B and LLaMA3.1-8B across categories of Gender, Religion, Profession and Region that proposes the India Context Sensitivity Index (ICSI) as a category-weighted fairness metric. The larger model shows an improvement of the aggregate bias score that is statistically significant (Mann-Whitney $p < 0.001$) a biased-response rate that is significantly lower ($\chi^2 = 51.92$, $p < 0.001$, 94.04% unbiased responses) at approximately double the inference latency. The breakdown of results by categories also shows that this overall gain is unevenly split: the 8B remains more sensitive on Gender-based prompts even as it improves on Religion, Profession and Region, underscoring the merit of reporting disaggregated, category-imbued fairness over a single number bias score for India deployed LLMs [9], [17], [23]. The future work will allow the scaling of the benchmark to be done to additional model sizes within the LLaMA family (for instance 3B, 70B) to instead fit a bias-versus-scale curve rather than a two-point comparison (14). Then, the extension of the category set to 'caste-adjacent' and intersectional categories (for instance gender × region) that are under-represented in this effort (18), (9). The next direction will be replacing the lexicon-based bias scorer with the LLM-as-judge scorer validated against human annotation, to measure scorer-induced bias in the evaluation pipeline itself (28). Finally, deploying the mitigation variants (few-shot, prompt-engineering) evaluated here as live runtime guardrails and measuring their effect on the ICSI in a closed deployment loop (29), (30).

REFERENCES

[1] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of 'bias' in NLP," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2020, pp. 5454–5476.

[2] I. O. Gallegos et al., "Bias and fairness in large language models: A survey," Computational Linguistics, vol. 50, no. 3, pp. 1097–1179, 2024.

[3] K. Khandelwal, M. Tonneau, A. M. Bean, H. R. Kirk, and S. A. Hale, "Casteist but not racist? Quantifying disparities in large language model bias between India and the West," arXiv preprint arXiv:2309.08573, 2023.

[4] A. Ahmad and P. Bhattacharyya, "Bias in language models: A survey," Tech. Rep., IIT Bombay CFILT, 2024.

[5] A. Urlana, C. V. Kumar, B. M. Garlapati, A. K. Singh, and R. Mishra, "No size fits all: The perils and pitfalls of leveraging LLMs vary with company size," in Proc. 31st Int. Conf. Comput. Linguistics (COLING), Industry Track, 2025, pp. 187–203.

[6] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "CrowS-Pairs: A challenge dataset for measuring social biases in masked language models," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 1953–1967.

[7] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2021, pp. 5356–5371.

[8] K. Khandelwal, M. Tonneau, A. M. Bean, H. R. Kirk, and S. A. Hale, "Indian-BhED: A dataset for measuring India-centric biases in large language models," in Proc. Int. Conf. Inf. Technol. Social Good (GoodIT), 2024, pp. 231–239.

[9] N. Sahoo, et al., "IndiBias: A benchmark dataset to measure social biases in language models for Indian context," in Proc. Conf. North American Chapter Assoc. Comput. Linguistics (NAACL), 2024, pp. 8786–8806.

[10] Meta AI, "The Llama 3 herd of models," arXiv preprint arXiv:2407.21783, 2024.

[11] A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.

[12] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2016, pp. 4349–4357.

[13] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in Proc. ACM Conf. Fairness, Accountability, Transparency (FAccT), 2021, pp. 610–623.

[14] Anonymous, "HALF: Harm-aware LLM fairness evaluation aligned with deployment," arXiv preprint arXiv:2510.12217, 2025.

[15] Anonymous, "Cross-platform evaluation of large language model safety in pediatric consultations: Evolution of adversarial robustness and the scale paradox," arXiv preprint arXiv:2601.09721, 2026.

[16] Anonymous, "CoBia: Constructed conversations can trigger otherwise concealed societal biases in LLMs," arXiv preprint arXiv:2510.09871, 2025.

[17] S. K. Dwivedi, S. Ghosh, and S. Dwivedi, "Breaking the bias: Gender fairness in LLMs using prompt engineering and in-context learning," Rupkatha J. Interdisciplinary Studies Humanities, vol. 15, no. 4, 2023.

[18] Anonymous, "DeCaste: Unveiling caste stereotypes in large language models through multi-dimensional bias analysis," arXiv preprint arXiv:2505.14971, 2025.

[19] J. A. Nawale, M. S. U. R. Khan, D. Janani, M. Gupta, D. Pruthi, and M. M. Khapra, “FairI tales: Evaluation of fairness in Indian contexts with a focus on bias and stereotypes,” arXiv preprint arXiv:2506.23111, 2025.

[20] Anonymous, “BharatBBQ: A multilingual bias benchmark for question answering in the Indian context,” arXiv preprint arXiv:2508.07090, 2025.

[21] Anonymous, “GenderBench: Evaluation suite for gender biases in LLMs,” arXiv preprint arXiv:2505.12054, 2025.

[22] Anonymous, “The effect of fine-tuning on language model toxicity,” arXiv preprint arXiv:2410.15821, 2024.

[23] T. Hofmann et al., “Explicitly unbiased large language models still form biased associations,” Proc. Nat. Acad. Sci. (PNAS), vol. 122, no. 8, 2025.

[24] Meta AI, “Llama 3.2: Lightweight and edge-ready models,” Meta AI Tech. Blog, 2024.

[25] S. Davidson, D. Bhattacharya, and I. Weber, “Racial bias in hate speech and abusive language detection datasets,” arXiv preprint arXiv:1905.12516, 2019.

[26] C. J. Hutto and E. Gilbert, “VADER: A parsimonious rule-based model for sentiment analysis of social media text,” in Proc. 8th Int. AAAI Conf. Weblogs and Social Media (ICWSM), 2014, pp. 216–225.

[27] A. Lees et al., “A new generation of perspective API: Efficient multilingual character-level transformers,” in Proc. 28th ACM SIGKDD Conf. Knowledge Discovery and Data Mining, 2022, pp. 3197–3207.

[28] K. K. B, P. Tiwari, A. C. Kumar, and A. Chandrabose, “Casteism in India, but not racism: A study of bias in word embeddings of Indian languages,” in Proc. LT4HALA Workshop, LREC, 2022, pp. 1–7.

[29] S. Kim et al., “Prometheus: Inducing fine-grained evaluation capability in language models,” arXiv preprint arXiv:2310.08491, 2023.

[30] A. Parrish et al., “BBQ: A hand-built bias benchmark for question answering,” in Findings of ACL 2022, pp. 2086–2105.

[31] M. F. Adilazuarda et al., “Towards measuring and modeling ‘culture’ in LLMs: A survey,” arXiv preprint arXiv:2403.15412, 2024.