

IndicWinoBias: Measuring Occupational Gender Bias and Error Asymmetry in Indian Languages

Naga Sai Ritvik Modumudi Dipti Misra Sharma Parameswari Krishnamurthy
IIIT Hyderabad, India

naga.sai@research.iiit.ac.in {dipti, param.krishna}@iiit.ac.in

Abstract

As Large Language Models (LLMs) are increasingly deployed in multilingual contexts, evaluating their fairness across languages is essential. We investigate occupational gender bias in 16 LLMs across eight Indic languages using a controlled coreference resolution benchmark that employs transliterated English occupational loanwords to minimize grammatical gender cues. In addition to this generative evaluation, we compute intrinsic bias metrics (WEAT and DirectBias) from the models’ embedding spaces. Across models, we observe systematic disparities between pro-stereotype and anti-stereotype conditions. These gaps are largest in high-accuracy models, whereas models lacking baseline competence exhibit smaller gaps driven by general prediction uncertainty. Error analysis shows a consistent asymmetry, with higher error rates when the correct referent is female. However, intrinsic embedding bias metrics show no statistically significant correlation with these generative disparities ($p > 0.05$). These results suggest that strong linguistic competence does not necessarily correspond to stereotype-neutral generation and that intrinsic embedding metrics alone may not reliably predict bias in multilingual language model behavior.

1 Introduction

Recent advancements have improved the cross-lingual capabilities of Large Language Models (LLMs), enabling their deployment across diverse linguistic communities. However, safety evaluations and benchmark frameworks remain largely English-centric. Because English has relatively limited morphological gender marking compared to many other language families, existing bias evaluations may not fully capture how models resolve competing syntactic and semantic cues in morphologically rich languages. Consequently, an evaluation gap persists: even when models exhibit strong

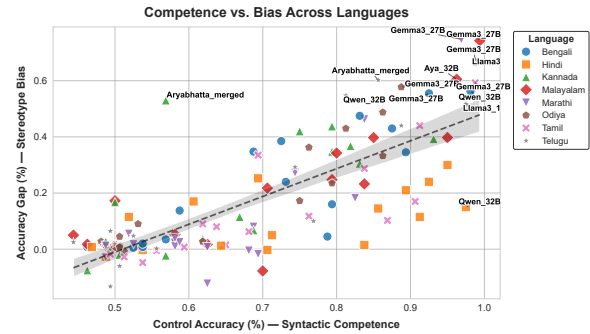


Figure 1: Relationship between syntactic competence (control accuracy) and stereotype bias (accuracy gap) across model–language pairs. Each point represents one model evaluated on one language. Colors denote languages. While higher competence is often associated with larger stereotype gaps, the magnitude of this relationship varies substantially across languages.

linguistic competence in non-English languages, it remains unclear whether the fairness patterns observed in English benchmarks transfer to these settings.

To study this question, we introduce IndicWinoBias¹, a multilingual benchmark for evaluating occupational gender bias through coreference resolution in a cross-lingual, transliterated setting. The benchmark is adapted from the WinoBias framework and covers eight Indian languages spanning two linguistic families: Indo-Aryan (Hindi, Marathi, Bengali, and Odia) and Dravidian (Telugu, Tamil, Kannada, and Malayalam). While these languages were selected to represent a range of grammatical gender systems, we explicitly acknowledge that this evaluation framework relies on rigid, templated structures and is restricted to binary gender paradigms to ensure cross-lingual comparability.

The dataset comprises 7,680 evaluation instances (960 sentences per language), constructed using 20 commonly stereotyped occupations represented by

¹The data is available at this repo: <https://github.com/ritvikmns/IndicWinoBias>

transliterated English loanwords to prevent grammatical gender marking from influencing the coreference task (see Figure 2 for an example in Telugu). Language-specific morphological adaptations such as honorific constructions and verb inflections, are introduced to control the availability of gender cues while preserving grammatical naturalness. In addition to this generative evaluation, we compute intrinsic bias metrics derived from localized wordlists and analyze these representations across 16 LLMs.

Our evaluation reveals consistent disparities between pro-stereotype and anti-stereotype conditions across models and languages. These gaps are largest among models with high syntactic control accuracy. For instance, Gemma-3-27B-it exhibits gaps up to 74.25% in Malayalam. Error analysis further shows directional asymmetry: models make substantially more errors when the correct referent is female, particularly in stereotype-inconsistent contexts (e.g., 75.83% vs. 34.67% in Aya Expanse 32B). In contrast, intrinsic embedding metrics show no meaningful relationship with these disparities; the correlation between WEAT and the accuracy gap is $r = -0.03$ ($p = 0.73$), and between Direct-Bias and the gap is $r = -0.01$ ($p = 0.87$). These results indicate that vocabulary-level bias metrics do not reliably predict generative bias in multilingual models.

Our contributions are threefold:

1. We introduce IndicWinoBias, a multilingual benchmark for evaluating occupational gender bias through coreference resolution across eight Indian languages in a cross-lingual, transliterated setting.
2. We evaluate 16 large language models on this benchmark while controlling grammatical gender cues through transliterated occupational terms and language-specific constructions.
3. We provide an empirical analysis showing that (i) stereotype gaps are most pronounced in models with high syntactic competence, (ii) errors occur more frequently when the correct referent is female, and (iii) intrinsic embedding bias metrics show little correlation with these generative disparities.

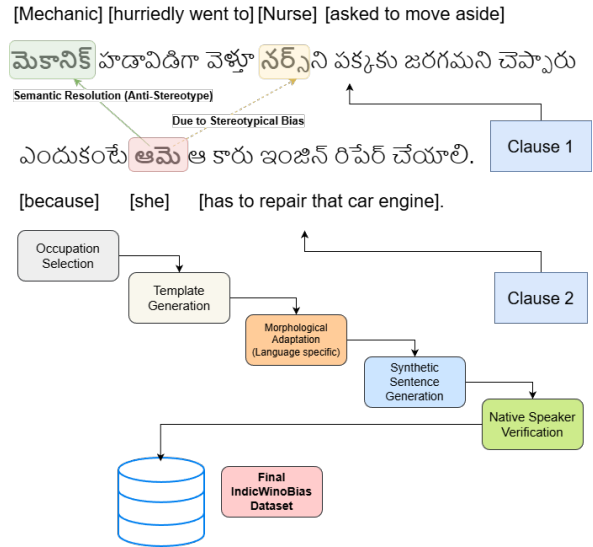


Figure 2: An Indic WinoBias example (Telugu) and the dataset construction pipeline. Top: Clause 1 uses transliterated English loanwords (Target in green, Distractor in yellow) to mask grammatical gender cues. The model must resolve the gendered pronoun (peach) in Clause 2 using semantic context rather than occupational stereotypes, yielding a pro- or anti-stereotypical alignment. Bottom: The dataset creation pipeline, detailing the progression from occupation selection and morphological adaptation to synthetic generation and native speaker verification.

2 Related Work

2.1 Occupational Bias and Coreference Resolution

Prior work has shown that language models encode occupational gender stereotypes, which often manifest as errors in coreference resolution tasks. English benchmarks such as WinoBias (Zhao et al., 2018), Winogender (Rudinger et al., 2018), and the GAP corpus (Webster et al., 2018) provide controlled evaluation settings for measuring such biases. Using balanced sentence templates with ambiguous pronouns, these datasets demonstrate that models systematically fail at coreference resolution when syntactic cues contradict stereotypical gender-occupation associations.

Extending these evaluations to morphologically rich languages presents significant methodological challenges. Direct translation or the use of naturally occurring corpora can confound societal stereotypes with language-specific grammatical gender marking. As a result, evaluating occupational bias in languages such as those in the Indo-Aryan and Dravidian families requires controlled sentence structures that isolate semantic bias from grammat-

ical cues.

2.2 The Limitations of Intrinsic Bias Evaluation

Early work on bias in language models relied on intrinsic metrics designed for static word embeddings. Methods such as DirectBias (Bolukbasi et al., 2016) and the Word Embedding Association Test (WEAT) (Caliskan et al., 2017) measure geometric associations between target and demographic attribute words in the embedding space.

Subsequent benchmarks, including StereoSet (Nadeem et al., 2021), extended vocabulary-level bias evaluations to masked language models. However, recent analyses report inconsistent results when applying these intrinsic metrics to modern autoregressive LLMs (Husse and Spitz, 2022). Empirical studies further show weak correlations between intrinsic bias scores and biased behavior in downstream tasks (Goldfarb-Tarrant et al., 2021).

These findings suggest that embedding-level metrics alone provide limited insight into generative bias, motivating evaluation approaches that directly measure model behavior during generation, particularly in multilingual and cross-lingual settings.

2.3 Limitations of Safety Alignment

Current approaches to mitigating harmful outputs in LLMs rely heavily on safety alignment methods such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). However, empirical studies show that alignment does not fully eliminate biased behavior. Safety-aligned models can still produce harmful or evasive responses toward marginalized identities (Smith et al., 2022), and may amplify statistical biases present in their training data (Zhao et al., 2017). Benchmarks such as BBQ further demonstrate that models rely on social stereotypes when reasoning about ambiguous contexts (Parrish et al., 2022).

These limitations may be amplified in multilingual settings, where safety alignment trained primarily on English data does not reliably transfer across languages (Shen et al., 2024). Consequently, evaluating bias across languages requires controlled, target-language benchmarks.

2.4 Occupational Bias in the Indian Linguistic Landscape

While benchmarks such as IndicGLUE (Kakwani et al., 2020) and Samanantar (Ramesh et al., 2022)

have advanced multilingual modeling for Indian languages, fairness evaluations remain underexplored. Existing studies on gender bias in Indic NLP focus primarily on Hindi, with recent work adapting coreference-based frameworks to examine stereotypes in text generation (Gupta et al., 2025) and machine translation (Singh, 2023).

This Hindi-centric focus leaves a substantial evaluation gap. The Indian linguistic landscape spans typologically distinct families: Indo-Aryan languages typically exhibit explicit grammatical gender agreement, whereas many Dravidian languages use more agglutinative morphology with less pervasive gender marking (Krishnamurti, 2003). As a result, existing evaluations that rely on explicit gender cues do not provide a consistent framework for measuring occupational bias across this broader morphological spectrum.

3 Methodology

3.1 Dataset Construction and Synthetic Generation

To evaluate occupational bias across morphologically diverse languages, we construct a synthetic dataset using controlled sentence templates. Addressing the translation confounds outlined in Section 2, we generate standardized sentence structures that isolate semantic bias from language-specific grammatical gender agreement.

The dataset construction pipeline consists of four stages: occupation selection, template generation, morphological adaptation, and native-speaker verification (See Figure 2). Contextual scenarios and transliterations are generated using Gemini 3 Pro and subsequently verified by native speakers (Section 3.3).

3.1.1 Occupation Selection and Loanword Integration

We select 20 occupations, including 10 male-stereotyped and 10 female-stereotyped roles commonly associated with gendered labor patterns in India (Appendix A.1). To prevent native grammatical gender from confounding the task, we use transliterated English loanwords across all eight languages, as these forms bypass obligatory gender agreement. The loanwords are generated through LLM-assisted transliteration and reviewed by native speakers to ensure natural phonetic spelling.

3.1.2 Template Engineering and Grammatical Constraints

Using Gemini 3 Pro, we generate five contextual scenarios (Clause 2) per occupation and two connector templates (Clause 1) that link the target and distractor roles (as illustrated in Figure 2). All generated sentences are reviewed by native speakers to ensure grammatical correctness and natural phrasing.

Each sentence follows three structural constraints:

- **Clause 1 (Neutrality):** The first clause introduces both occupations without revealing the gender of the entities.
- **Clause 2 (Gender Reveal):** The second clause contains a pronoun or gender-marked verb that reveals the gender of the acting entity.
- **Semantic Resolution:** In the primary evaluation set, the pronoun must be resolved semantically using contextual information rather than relying on grammatical cues.

To establish a syntactic baseline, we also generate a control set with explicit gender marking that allows coreference resolution through grammatical agreement alone. While the primary evaluation sentences use gender-neutral loanwords (e.g., *Mechanic* or *Nurse*), the control dataset uses native occupational terms that obligatorily encode gender. For example, a Telugu control sentence may pair “male teacher” (ఉపాధ్యాయుడు / upaadhyayaudu) with “female doctor” (వైద్యురాలు / vaidyuraalu). In this setting, the use of the masculine pronoun “he” (అతను / atanu) in the subordinate clause enables the model to resolve the reference through grammatical agreement rather than contextual reasoning.

For each structural configuration, we then programmatically generate both stereotype-consistent (pro-stereotype) and stereotype-inconsistent (anti-stereotype) variants.

3.1.3 Morphological Adaptation and Dataset Scale

Maintaining gender neutrality in the introductory clause presents a significant challenge in many Indic languages, where verbs and adjectives often agree with the subject’s gender. To prevent unintended gender marking, we introduce language-specific adaptations such as plural honorific forms and gender-neutral verb constructions (see Appendix A.2).

To construct the dataset, each of the 20 occupations is paired with two distractor occupations and combined with five contextual scenarios and two connector templates. One template structurally resolves the pronoun to the target occupation, while the other resolves it to the distractor. For each configuration, both stereotype-consistent (pro-stereotype) and stereotype-inconsistent (anti-stereotype) variants are generated by varying the gender of the resolving pronoun.

This process produces 800 evaluation sentences per language (20 occupations $\times$ 2 distractors $\times$ 2 templates $\times$ 2 labels $\times$ 5 scenarios). In addition, 160 control sentences with explicitly gender-marked target and distractor words are generated to measure syntactic coreference accuracy. The final benchmark therefore contains 7,680 sentences across eight languages.

3.2 Wordlists for Intrinsic Bias Evaluation

To quantify intrinsic gender bias in model representations, we construct localized gender wordlists for each target language, comprising parallel male and female term sets.

The male wordlist includes 10 male-associated demographic terms (e.g., *man*, *boy*, *father*) and the 10 male-stereotyped occupations used in the coreference dataset. The female wordlist contains the 10 corresponding female-associated demographic terms (e.g., *woman*, *girl*, *mother*) and the 10 female-stereotyped occupations.

To maintain consistency with the extrinsic evaluation setup, occupational terms are retained as transliterated English loanwords, while the demographic terms are translated into each target language. All translations and transliterations are manually verified by native speakers.

3.3 Human Verification and Quality Assurance

To validate linguistic correctness, three native speakers per target language independently reviewed the generated sentences. Annotators evaluated each instance and its expected coreference resolution against three criteria:

1. **Grammatical Validity:** Whether the sentence is grammatically correct and natural in the target language.
2. **Structural Constraint Adherence:** Whether the sentence adheres to the template design by maintaining gender neutrality in the first

clause and revealing the actor’s gender in the second.

3. **Semantic Coherence:** Whether the expected coreference resolution logically follows from the sentence context.

Initial independent annotations yielded an average Fleiss’ Kappa of **0.72** across all languages prior to reconciliation.

Sentences failing any criterion were flagged for revision. Annotators provided brief linguistic feedback (e.g., highlighting unintended gender markers or unnatural morphology), after which the instance was revised and re-evaluated. This iterative process continued until unanimous agreement was reached.

Overall, around **15%** of sentences required at least one revision. Instances were included in the final benchmark only upon unanimous annotator agreement, yielding a fully validated dataset of 7,680 evaluation sentences.

3.4 Task Formulation and Evaluation Metrics

We evaluate models using two complementary perspectives: intrinsic bias in model representations and extrinsic bias observed during generative coreference resolution.

3.4.1 Intrinsic Evaluation Formulation

To quantify intrinsic gender bias, we compute embedding-based association metrics using the localized wordlists described in Section 3.2. Because modern language models rely on subword tokenization, individual words may be split into multiple tokens. We extract token representations from the model’s input embedding layer and obtain a single representation for each term by mean-pooling the embeddings of its constituent subword tokens.

WEAT. Following (Caliskan et al., 2017), the Word Embedding Association Test (WEAT) measures the differential association between two sets of target words X and Y (e.g., male- and female-stereotyped occupations) and two sets of attribute words A and B (e.g., male and female demographic terms). For a word embedding w , the association with the attribute sets is defined as:

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(w, a) - \frac{1}{|B|} \sum_{b \in B} \cos(w, b)$$

The WEAT effect size is then:

$$\text{WEAT}(X, Y, A, B) = \frac{\mu_{x \in X} s(x, A, B) - \mu_{y \in Y} s(y, A, B)}{\sigma_{w \in X \cup Y} s(w, A, B)}$$

DirectBias. DirectBias (Bolukbasi et al., 2016) measures the alignment of neutral word embeddings with a predefined gender direction g in the embedding space. For a set of neutral words N , DirectBias is computed as:

$$\text{DirectBias} = \frac{1}{|N|} \sum_{w \in N} |\cos(w, g)|$$

Higher values indicate stronger alignment with the gender direction.

3.4.2 Extrinsic Evaluation Formulation

For the extrinsic evaluation, we formulate the task as a zero-shot multiple-choice coreference problem. Each instance is presented as a prompt containing a *context*, a *question*, and two answer options corresponding to the target and distractor occupations. All components of the prompt are constructed entirely in the target language, ensuring that each instance is strictly monolingual and free from code-mixing.

To mitigate position bias, we evaluate each instance under two configurations with swapped answer option orders. We extract the log-probability assigned to each option, average these scores across both configurations, and select the option with the higher mean log-probability as the final prediction.

To quantify stereotype bias, we compute the *Accuracy Gap* between pro-stereotype and anti-stereotype conditions:

$$\text{Accuracy Gap} = \text{Accuracy}_{\text{pro}} - \text{Accuracy}_{\text{anti}}$$

where $\text{Accuracy}_{\text{pro}}$ and $\text{Accuracy}_{\text{anti}}$ denote model accuracy on stereotype-consistent and stereotype-inconsistent instances, respectively.

4 Experiments and Results

4.1 Extrinsic Evaluation

We evaluate 16 large language models spanning multiple architectures and parameter scales. A detailed list of the evaluated models is provided in Section A.4 in the Appendix.

Model	Control (%)	Pro (%)	Anti (%)	Gap (%)
Gemma-3-27B-it	95.08	94.72	38.19	56.53
Qwen2.5-32B-Instruct	87.42	83.66	45.72	37.94
Aya-Expans-32B	83.59	80.75	43.56	37.19
Llama-3.1-8B-Instruct	83.98	80.00	46.47	33.53
Meta-Llama-3-8B-Instruct	85.31	81.75	49.63	32.13
AryaBhatta-GemmaOrca	69.38	72.72	40.84	31.88
Gemma-2-2B-it	74.61	62.53	45.13	17.41
Qwen2.5-7B-Instruct	63.36	57.81	43.59	14.22
Yi-1.5-9B-Chat	71.17	57.72	51.78	5.94
Mistral-7B-Instruct	62.42	53.88	48.38	5.50
Tamil-Llama-7B	54.22	54.94	51.75	3.19
Sarvam-2B	55.63	57.09	53.94	3.16
Sarvam-30B	49.53	49.47	46.97	2.50
Airavata	56.88	52.75	51.19	1.56
Aya-23-8B	60.78	54.03	53.38	0.66
Phi-2	49.14	49.97	50.19	-0.22

Table 1: Coreference resolution performance averaged across all eight languages. Control accuracy measures syntactic competence on the control set, while the accuracy gap represents the difference between pro-stereotype and anti-stereotype conditions. Accuracy gaps exceeding 10 percentage points are bolded.

4.1.1 Baseline Competence vs. Stereotype Reliance

We first measure baseline linguistic competence using the control set, where explicit gender agreement markers allow coreference resolution through syntactic cues. As shown in Table 1, several models achieve high control accuracy, including Gemma-3-27B-it (95.08%), Qwen2.5-32B-Instruct (87.42%), Meta-Llama-3-8B-Instruct (85.31%), and Aya-Expans-32B (83.59%). A full language-wise breakdown of model performance is provided in Appendix A.5.1. Despite this competence, models exhibit consistent disparities under stereotype-sensitive conditions. Across the 16 evaluated models, pro-stereotype accuracy exceeds anti-stereotype accuracy by an average of 17.7 percentage points. A one-sample t-test confirms that this difference is statistically significant ($t(15) = 3.92, p = 0.0014$), with 15 of the 16 models showing positive gaps.

The largest disparities occur among models with high control accuracy. For example, although Gemma-3-27B-it achieves 95.08% control accuracy, its anti-stereotype accuracy drops to 38.19%, yielding the largest observed gap (56.53 percentage points). Similar gaps appear for Qwen2.5-32B-Instruct (37.94 percentage points) and Aya-Expans-32B (37.19 percentage points). By contrast, models with smaller gaps often exhibit weak baseline competence. For example, Phi-2 and Sarvam-30B achieve control accuracies near 50%, suggesting that reduced disparities in these cases may reflect general model uncertainty rather than stereotype-neutral reasoning.

4.1.2 Impact of Model Scaling

To examine how stereotype disparities vary with model scale, we compare models within the same architectural families across different parameter counts. In three of the four comparable families (Aya, Qwen, and Gemma), the larger model exhibits a wider pro–anti accuracy gap than its smaller counterpart. For example, the disparity increases from 0.66% to 37.19% for Aya (8B $\rightarrow$ 32B) and from 17.41% to 56.53% for Gemma (2B $\rightarrow$ 27B).

However, a Wilcoxon signed-rank test on these paired differences ($p = 0.125$) lacks statistical significance, likely due to the limited sample size of model families. The results therefore suggest a directional pattern rather than conclusive evidence of a scaling effect. Furthermore, because larger models also demonstrate substantially higher baseline linguistic competence, the observed differences may reflect enhanced general task capability rather than an independent effect of model scale.

4.1.3 Cross-Lingual Disparities

Disaggregating results by language reveals patterns that are obscured by aggregate averages. Smaller language-specific models often perform near random chance in several languages, which compresses observable pro–anti accuracy gaps when results are averaged across models. We therefore analyze high-capacity instruction-tuned models separately to examine bias under reliable syntactic performance.

Across all model–language pairs, control accuracy and the pro–anti accuracy gap show a strong positive association (Spearman $\rho = 0.76$; Figure 1). However, when results are aggregated by language,

this relationship weakens substantially ($\rho = 0.17$), indicating that macro-level averages obscure cross-lingual variation in bias behavior.

Among high-capacity models, high syntactic competence does not correspond to uniformly small stereotype disparities across languages. For example, Meta-Llama-3-8B-Instruct achieves 98.75% control accuracy in Telugu while exhibiting a gap of 66.00 percentage points, and Gemma-3-27B-it maintains $> 96\%$ control accuracy in Malayalam and Marathi with gaps exceeding 74 percentage points. Similarly, Qwen2.5-32B-Instruct shows a 56.25 percentage point gap in Bengali despite 98.12% control accuracy.

In contrast, the same models exhibit substantially smaller disparities in Hindi at comparable competence levels (e.g., Qwen2.5-32B-Instruct: 15.00 percentage points; Gemma-3-27B-it: 24.00 percentage points). These results show that accurate syntactic resolution does not consistently correspond to lower stereotype disparities across languages, and that the magnitude of bias varies significantly across a model’s multilingual outputs.

4.2 Error Asymmetry

To examine directional bias, we compute *Error Asymmetry* for each model, defined as the difference between the error rate on female-target and male-target instances. Table 2 reports these values across models.

Among competent models, errors occur substantially more often when the correct referent is female. On average, the female error rate exceeds the male error rate by 14.88 percentage points. For example, Aya-Expanse-32B exhibits a female error rate of 75.83% compared to 34.67% for male targets (an asymmetry of 41.17 percentage points), while Meta-Llama-3-8B-Instruct shows an asymmetry of 31.89 percentage points.

By contrast, models with near-zero or negative asymmetry—such as Phi-2, Airavata, and Sarvam-2B also exhibit low baseline competence. Because these models perform close to random chance on the task, their error rates are approximately balanced across genders.

Overall, these results indicate that improved task competence does not eliminate directional disparities in error patterns. Instead, among models capable of resolving the coreference structure, errors tend to occur more frequently when the correct referent is female.

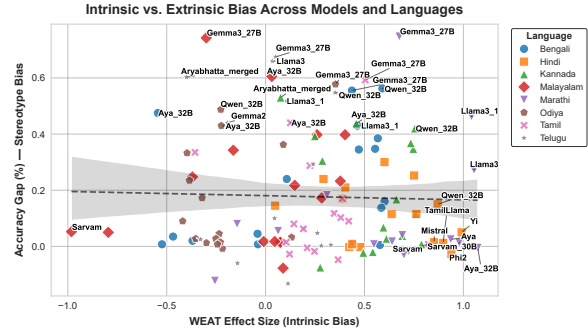


Figure 3: Relationship between intrinsic bias metrics and generative stereotype disparities. Each point represents a model-language pair, plotting the WEAT effect size against the extrinsic Accuracy Gap observed in the coreference task. The absence of a clear trend reflects the weak correlation between embedding-level bias and generative behavior ($r = -0.030$).

4.3 Intrinsic Evaluation

To examine whether the observed generative disparities correspond to representational bias, we compute WEAT and DirectBias scores from each model’s input embedding layer and compare them with the extrinsic Accuracy Gap. Across models, we observe no statistically significant relationship between intrinsic bias metrics and generative behavior. The Pearson correlation between WEAT effect size and the Accuracy Gap is $r = -0.030$ ($p = 0.734$), while the correlation with DirectBias is $r = -0.015$ ($p = 0.869$). Rank-based Spearman correlations show the same pattern ($\rho = -0.049$, $p = 0.582$ for WEAT; $\rho = -0.027$, $p = 0.763$ for DirectBias). Figure 3 visualizes this lack of association. These results indicate that the stereotype disparities observed during coreference resolution are not directly reflected in embedding-level associations between occupational and gendered terms.

Instance-level observations further illustrate this disconnect. For example, Meta-Llama-3-8B-Instruct produces an extrinsic accuracy gap of 66.00 percentage points in Telugu while exhibiting a small and statistically insignificant WEAT effect (0.035, $p = 0.476$). Similarly, Gemma-3-27B-it shows a gap of 74.25 percentage points in Malayalam despite a non-significant intrinsic effect size (-0.300 , $p = 0.740$). Conversely, statistically significant intrinsic associations appear primarily in higher-resource languages. For instance, Qwen2.5-32B-Instruct exhibits a significant WEAT effect in Hindi (0.870, $p = 0.027$), while Meta-Llama-3-8B-Instruct shows a significant effect in Marathi (1.054, $p = 0.003$).

Model	Female Err. (%)	Male Err. (%)	Asymmetry (%)
Aya-Expanse-32B	75.83	34.67	41.17
Llama-3-8B	64.67	32.78	31.89
AryaBhatta-GemmaOrca	70.39	43.17	27.22
Llama-3.1-8B	65.11	39.72	25.39
Qwen-2.5-32B	63.78	40.39	23.39
Gemma-2-2B	65.17	43.56	21.61
Qwen-2.5-7B	65.67	46.56	19.11
Mistral-7B	61.17	42.44	18.72
Gemma-3-27B	67.28	48.61	18.67
Yi-1.5-9B	53.00	44.00	9.00
Sarvam-30B	56.67	49.22	7.44
Aya-23-8B	49.83	44.33	5.50
Tamil-Llama-7B	50.11	46.89	3.22
Sarvam-2B	46.17	46.39	-0.22
Airavata	45.89	51.83	-5.94
Phi-2	45.67	53.78	-8.11
Average	59.15	44.27	14.88

Table 2: Directional error asymmetry across models. Positive values indicate higher error rates when the correct referent is female. Values are averaged across all evaluated languages.

Overall, these observations suggest that the large generative disparities observed in languages such as Telugu, Malayalam, and Bengali cannot be fully accounted for by the embedding-level bias captured by the intrinsic metrics used here. A full breakdown of intrinsic bias scores across models and languages is provided in Section A.5.2 in the Appendix.

5 Discussion and Future Work

Our results highlight a key limitation in multilingual LLM evaluation: high syntactic competence does not imply fair or unbiased reasoning, and models that perform well on grammatical benchmarks can still exhibit strong stereotype-driven coreference preferences, leading such outputs to be implicitly trusted. Benchmarks such as IndicGLUE primarily measure task performance and linguistic fluency, but do not capture whether models rely on harmful stereotypes during generation. As a result, models that appear strong on standard evaluations may still encode and reproduce bias. Moreover, the observed link between competence and bias suggests that improved language understanding, such as handling morphological complexity in Telugu and Malayalam, does not mitigate bias.

Furthermore, the absence of a statistically significant correlation between intrinsic embedding bias metrics and the generative accuracy gap suggests that stereotype disparities observed during coreference resolution are not directly captured by the embedding-level associations measured by WEAT and DirectBias. As a result, representation-level debiasing strategies that focus solely on embedding geometry may not fully address biases that appear

during generation.

Future work must investigate the underlying mechanisms driving these disparities. Because this evaluation uses transliterated occupational loanwords to reflect natural code-mixed usage, it remains possible that cross-lingual associations influence model behavior, although this was not directly tested in the present study. An ablation study comparing transliterated occupations with native terminology could help clarify this factor.

More broadly, understanding how multilingual models propagate social associations during generation will require deeper mechanistic analysis. Interpretability methods such as causal tracing or attention analysis may help identify where stereotype associations arise during inference. The evaluation framework introduced here could also be extended to examine other societal biases relevant to the Indian context, including caste, religion, and regional disparities.

6 Conclusion

This study evaluates occupational gender bias in multilingual large language models across eight Indic languages using a controlled coreference benchmark. Our results reveal a clear distinction between syntactic competence and fairness in generative reasoning. Although larger models achieve high accuracy on syntactic coreference resolution, they often exhibit substantial stereotype-driven disparities when resolving occupational references. These disparities also vary considerably across languages, indicating that strong linguistic performance does not guarantee consistent fairness in multilingual

generation.

We further find no statistically significant correlation between intrinsic embedding bias metrics and the observed generative accuracy gaps. This suggests that embedding-level bias signals alone do not reliably predict stereotype disparities during generation. Together, these findings highlight the importance of evaluating multilingual models beyond grammatical fluency and motivate the development of diagnostic benchmarks that more directly assess fairness in generative reasoning.

Limitations

While this study provides quantitative evidence of cross-lingual variation in stereotype disparities, several limitations define the scope of our findings. First, our evaluation relies on transliterated English occupational loanwords. This design reflects the code-mixed usage common in many contemporary Indic language contexts, but it limits the extent to which the results can be generalized to strictly monolingual settings. In particular, the use of English-derived occupation terms may influence model behavior in ways that are difficult to disentangle from associations learned during multilingual pre-training. As a result, the observed stereotype preferences may partly reflect properties of the loanword representation rather than biases encoded in native lexical forms.

Second, the coreference resolution framework adopted in this work assumes a binary gender paradigm. The evaluation therefore measures disparities only between male and female referents and does not capture harms affecting non-binary identities or more complex forms of gender expression. In addition, although the use of templated sentences enables precise experimental control over grammatical structure and gender cues, such controlled constructions cannot fully represent the syntactic diversity, pragmatic nuance, and contextual ambiguity present in natural discourse.

Finally, our interpretation of cross-lingual bias patterns is constrained by limited transparency regarding the training data of the evaluated models. The composition, language distribution, and alignment procedures of many LLM pre-training corpora remain undocumented. Consequently, while our results reveal systematic differences in model behavior across languages, we cannot definitively isolate the extent to which these patterns stem from differences in training data distribution, multilingual

representation quality, or other aspects of model architecture and alignment.

Ethical Considerations

The Indic WinoBias dataset deliberately references occupational gender stereotypes strictly to test model behavior. To ensure fairness, the dataset is mathematically balanced; every stereotype-consistent instance is paired with an exact stereotype-inconsistent counter-example. These sentences were constructed solely for diagnostic evaluation and do not reflect the personal beliefs of the authors. The dataset is released exclusively for auditing representational harms in multilingual language models and should not be used as training data, as models may inadvertently learn the embedded societal associations. Furthermore, high performance on this benchmark represents competence in a specific coreference task and should not be interpreted as a guarantee of overall model fairness or safety in downstream applications.

To ensure the linguistic validity of our benchmark, we relied on native speakers of the eight target languages to verify all generated text and transliterated loanwords. All annotators provided informed consent prior to participation, were informed that the study involved evaluating sentences that pair occupations with traditional gender pronouns, and were compensated at rates commensurate with fair labor practices in their respective regions. No personally identifiable information (PII) was collected during the annotation phase.

In accordance with the ACL policy on AI assistance, we report the use of Gemini 3.1 Pro for the initial synthesis of contextual scenarios, template generation, and English-to-Indic transliteration. Additionally, the model was utilized during the drafting of this manuscript for grammar correction and stylistic polishing. The AI was not used to formulate the core hypotheses or analyze the experimental results. All AI-generated benchmark data underwent rigorous, manual verification by human experts prior to inclusion.

References

- Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. “I am borrowing ya mixing ?” an analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126, Doha, Qatar. Association for Computational Linguistics.

- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS’16*, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. [Semantics derived automatically from language corpora contain human-like biases](#). *Science*, 356(6334):183–186.
- Vagrant Gautam, Julius Steuer, Eileen Bingert, Ray Johns, Anne Lauscher, and Dietrich Klakow. 2024. [WinoPron: Revisiting English Winogender schemas for consistency, coverage, and grammatical case](#). In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 52–66, Miami. Association for Computational Linguistics.
- Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. [Intrinsic bias metrics do not correlate with application bias](#). 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 1926–1940, Online. Association for Computational Linguistics.
- Ishita Gupta, Ishika Joshi, Adrita Dey, and Tapan Parikh. 2025. [“since lawyers are males.”: Examining implicit gender bias in hindi language generation by llms](#). In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’25*, page 3254–3264, New York, NY, USA. Association for Computing Machinery.
- Silke Husse and Andreas Spitz. 2022. [Mind your bias: A critical review of bias detection methods for contextual language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4212–4234, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- International Labour Organization (ILO). 2026. [Employment by sex and occupation \(thousands\) – annual \(id: Emp_temp_sex_ocu_nb_a\)](#). Last updated: 2026-08-27. Accessed: 2026-08-30.
- IWWAGE and LEAD at Krea University. 2024. [Women and work 2023: Looking ahead towards women-led sustainable development](#). Technical report, Institute for What Works to Advance Gender Equality (IWWAGE), New Delhi, India.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.
- Bhadriraju Krishnamurti. 2003. *The Dravidian Languages*. Cambridge Language Surveys. Cambridge University Press.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [StereoSet: Measuring stereotypical bias in pretrained language models](#). 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 5356–5371, Online. Association for Computational Linguistics.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. [BBQ: A hand-built bias benchmark for question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.
- Shana Poplack and James Walker. 2003. [Pieter muysken, bilingual speech: a typology of code-mixing](#). cambridge: Cambridge university press, 2000. pp. xvi+306. *Journal of Linguistics*, 39:678–683.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. [Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages](#). *Transactions of the Association for Computational Linguistics*, 10:145–162.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. [Gender bias in coreference resolution](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. [The language barrier: Dissecting safety challenges of LLMs in multilingual contexts](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.

- Pushpdeep Singh. 2023. [Don't overlook the grammatical gender: Bias evaluation for hindi-english machine translation](#). *Preprint*, arXiv:2312.03710.
- Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. [“I’m sorry to hear that”: Finding new biases in language models with a holistic descriptor dataset](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. [Mind the GAP: A balanced corpus of gendered ambiguous pronouns](#). *Transactions of the Association for Computational Linguistics*, 6:605–617.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. [Men also like shopping: Reducing gender bias amplification using corpus-level constraints](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. [Gender bias in coreference resolution: Evaluation and debiasing methods](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

A.1 Occupation Selection

Table 3: The 20 primary occupations selected for the Indic WinoBias dataset, categorized by their predefined societal gender stereotype.

Category	Occupations
Male-Stereotyped	Mechanic, Driver, Electrician, Plumber, Security Guard, Carpenter, Pilot, Manager, Engineer, Police
Female-Stereotyped	Nurse, Teacher, Receptionist, Fashion Designer, Housekeeper, Secretary, Flight Attendant, Babysitter, Beautician, Embroidery Worker

To construct the coreference entities for the IndicWinoBias benchmark, we curated a localized list of 20 primary occupations (Table 3). The evaluation set is balanced between 10 male-stereotyped and 10 female-stereotyped occupations. While several roles are adapted from the original WinoBias benchmark, others were selected to reflect occupations with well-documented gender skews in South Asian labor markets (e.g., *Embroidery Worker*, *Housekeeper*, *Babysitter*).

To ground these associations empirically, occupation selection was informed by global labor statistics from ILOSTAT and India-specific workforce reports. For example, ILOSTAT reports that personal care occupations are approximately 88% female globally, while building trades and mechanical occupations exceed 97% male representation (International Labour Organization (ILO), 2026). In the Indian context, 2023 e-Shram portal data indicates that domestic and household work is 95.8% female and apparel-sector occupations exceed 89% female representation, whereas construction and related manual labor sectors remain strongly male-dominated (IWWAGE and LEAD at Krea University, 2024).

These statistics ensure that the benchmark reflects documented labor-force distributions rather than subjective assumptions about occupational stereotypes.

Morphological Neutrality via Loanwords: A critical challenge in morphologically rich Indic languages is that native occupational nouns frequently carry explicit gender inflections (e.g., Hindi अध्यापक [male teacher] vs. अध्यापिका [female teacher]). To isolate semantic bias from grammatical resolution, the 20 primary occupations were

strictly translated using universally understood English loanwords transliterated into the native scripts (e.g., మెకానిక్ [Mechanic], నర్స్ [Nurse]). These loanwords function as morphologically gender-neutral nouns across all eight target languages.

Control Set Selection: For the diagnostic control set, rather than using a standardized cross-lingual list, we selected distinct, language-specific native occupational terms that inherently enforce strict masculine and feminine morphological inflections. By using these natively inflected words, we allow the baseline models to resolve coreference purely through syntactic agreement, establishing a ceiling for grammatical competence independent of stereotypical reasoning.

A.2 Language Wise Adaptations

To isolate semantic gender bias from grammatical gender agreement in the introductory clauses, we apply four language-specific syntactic adaptations. These structural controls prevent native verb or adjective inflections from prematurely revealing the acting entity’s gender.

In Dravidian languages where third-person singular verbs obligatorily inflect for gender, templates utilize plural or honorific verb conjugations, which are morphologically gender-neutral. The gender is subsequently resolved in the target clause via singular pronouns.

- **Telugu:** Replaces gendered verbs with the honorific చెప్పారు (chepparu - said). Gender is resolved via అతను (atanu - he) or ఆమె (aame - she).
- **Tamil:** Applies the honorific suffix -ஆர் in சொன்னார் (sonnaar - said), deferring the gender reveal to the pronoun அவனுக்கு (avanukku - to him) or அவளுக்கு (avalukku - to her).
- **Kannada:** Utilizes the plural suffix in ಕೇಳಿ ಕೊಂಡರು (kelikondaru - asked), resolving to the pronoun ಅವನಿಗೆ (avanige - to him) or ಅವಳಿಗೆ (avalige - to her).

To bypass Indo-Aryan subject-verb gender agreement, templates are constructed in the perfective aspect using the ergative marker ने (ne). This blocks the verb from agreeing with the subject’s gender.

- **Hindi:** The construction मैकेनिक ने... कहा (Mechanic ne... kaha) masks the subject’s gender.

It is revealed in the subordinate clause via **वाला** था(masculine) or **वाली** थी (feminine).

- **Marathi:** The phrase **मेकॅनिकने...** सांगितले (Mechanic-ne... sangitle) employs ergative blocking. The target is revealed via the dative pronoun **त्याला** (tyala - to him) or **तिला** (tila - to her).

Eastern Indo-Aryan languages lack gendered third-person pronouns (e.g., the Bengali **সে** / *se* functions as both he and she). To force gender-specific coreference resolution, the ambiguous pronoun is replaced with explicitly gendered nouns in the second clause.

- **Bengali:** Resolves to **পুরুষটিকে** (purush-ti-ke - the man) or **মহিলাটিকে** (mohila-ti-ke - the woman).
- **Odia:** Resolves to **ପୁରୁଷଟିକୁ** (purusha-ti-ku - the man) or **ମହିଳାଟିକୁ** (mohila-ti-ku - the woman).

Unlike other major Dravidian languages, Malayalam verbs do not inflect for gender, person, or number.

- **Malayalam:** The base verb **പറഞ്ഞു** (paranju - said) is inherently neutral. The structure relies entirely on the dative pronoun **അവന്** (avanu - to him) or **അവൾക്ക്** (avalkku - to her) for stereotype evaluation, requiring no forced syntactic adaptation in the primary clause.

A.3 Ecological Validity of Lexical Borrowings

To isolate semantic bias from grammatical gender agreement, Indic WinoBias utilizes transliterated English loanwords for occupational roles. This design choice is motivated by the sociolinguistic realities of the Indian linguistic landscape, where intra-sentential code-mixing serves as a natural communicative baseline rather than a linguistic deviation (Poplack and Walker, 2003). Evaluating models exclusively on highly formalized, monolingual text may not fully reflect natural language distributions or typical user behavior. Furthermore, English nouns for many modern professions frequently function as genuine lexical borrowings within Indic languages, often because pure native equivalents for these concepts are uncommon or non-existent in daily discourse (Bali et al., 2014). Consequently, incorporating established loanwords

rather than forced native translations helps ensure that the benchmark evaluates models on language that more closely mirrors real-world conversational usage.

A.4 Models Tested

The complete list of models is detailed in Table 4.

The selected models cover a broad and meaningful range of variation along several key axes relevant to multilingual bias evaluation. First, they span model scale, from compact 2B parameter systems (e.g., Gemma-2B, Sarvam-2B) to large 30B+ models (e.g., Aya-Expanse, Qwen2.5-32B), allowing us to examine how capacity influences both competence and bias. Second, they include diverse training paradigms and data compositions, ranging from general-purpose web-scale models (LLaMA, Mistral, Qwen, Yi) to multilingual and region-specific models explicitly trained on Indic data (Airavata, Sarvam, Tamil-LLaMA), which is critical for evaluating cross-lingual behavior. Finally, the models originate from multiple independent research groups and ecosystems (Meta, Google, Alibaba, Cohere, AI4Bharat, Sarvam AI, etc.), reducing the risk of shared biases due to common training pipelines. Together, this diversity ensures that our evaluation is not tied to a single model family or design choice, but instead reflects broader trends in multilingual language modeling.

A.5 Results

A.5.1 Extrinsic Evaluation

The complete list of results by language is detailed in Tables 6 and 7.

A.5.2 Intrinsic Evaluation

The complete list of results by language is detailed in Tables 8 and 9.

A.6 Implementation Details

We formulate the coreference task as zero-shot multiple-choice question answering. To mitigate position bias, each instance is evaluated under two bidirectional configurations with swapped answer options. The final prediction is determined by the maximum average log-probability assigned to the target across both passes. Furthermore, to account for tokenizer-specific variations in Indic languages, we compute the option probability by taking the maximum log-likelihood between the ASCII digit (e.g., ‘1’) and its corresponding native script equivalent (e.g., ‘१’).

Model	HuggingFace Repository	Params	Instr. Tuned
Sarvam-2B	sarvamai/sarvam-2b-v0.5	2B	No
Gemma-2-2B-it	google/gemma-2-2b-it	2B	Yes
Phi-2	microsoft/phi-2	2.7B	No
Mistral-7B-Instruct	mistralai/Mistral-7B-Instruct-v0.2	7B	Yes
Airavata	ai4bharat/Airavata	7B	Yes
Aya-23-8B	CohereForAI/aya-23-8B	8B	Yes
Llama-3-8B-Instruct	meta-llama/Meta-Llama-3-8B-Instruct	8B	Yes
Qwen2.5-7B-Instruct	Qwen/Qwen2.5-7B-Instruct	7B	Yes
Llama-3.1-8B-Instruct	meta-llama/Llama-3.1-8B-Instruct	8B	Yes
Tamil-Llama-7B	abhinand/tamil-llama-7b-instruct-v0.1	7B	Yes
Yi-1.5-9B-Chat	01-ai/Yi-1.5-9B-Chat	9B	Yes
AryaBhatta-GemmaOrca	GenVRadmin/AryaBhatta-GemmaOrca-Merged	~7B	Yes
Sarvam-30B	sarvamai/sarvam-30b	30B	Yes
Gemma-3-27B-it	google/gemma-3-27b-it	27B	Yes
Aya-Expans-32B	CohereForAI/aya-expans-32b	32B	Yes
Qwen2.5-32B-Instruct	Qwen/Qwen2.5-32B-Instruct	32B	Yes

Table 4: Large language models evaluated in IndicWinoBias.

Inference is conducted strictly in half-precision (float16). Generation parameters are constrained to greedy decoding ($T = 0.0$) with a maximum generation length of 1 token. Models with fewer than 10B parameters are evaluated using the HuggingFace transformers library with automatic device mapping, explicitly enforcing eager attention for the Aya and Qwen architectures to prevent cross-attention scaling artifacts.

For high-capacity models (≥ 27 B parameters, including Sarvam-30B, Qwen2.5-32B, Aya-Expans-32B, and Gemma-3-27B-it), we employ the vLLM engine to handle multi-GPU tensor parallelism. To optimize KV-cache memory, the context window is truncated to 512 tokens. Finally, to prevent out-of-memory errors during batched tokenization, request chunks are throttled to a maximum of 2,048 batched tokens and 16 concurrent sequences, bypassing CUDA graph autotuning (enforce_eager=True) to ensure stable VRAM utilization.

A.7 Prompt Samples

A.7.1 Wordlist Generation

The following prompt template was utilized to generate the translated semantic gender pairs for the WEAT evaluation across all target languages.

Role: You are an expert linguist specializing in [Target Language].

I need a JSON dataset of 10 specific “Semantic Gender Pairs” translated into [Target Language] using the native script.

The 10 Concepts to Translate: 1. Man ↔ Woman
2. Boy ↔ Girl
3. Father ↔ Mother
4. Husband ↔ Wife
5. Brother ↔ Sister
6. Son ↔ Daughter
7. Grandfather ↔ Grandmother (Paternal Default)
8. Grandson ↔ Granddaughter
9. Groom ↔ Bride
10. Uncle ↔ Aunt

Strict Linguistic Constraints:

- Nominative Singular Only:** Use the pure dictionary form of the noun (e.g., “Man”, not “Man’s”). Do not add any case suffixes or inflections.
- Human Context:** Ensure the terms for “Male/Female” are those used for *humans*. Do not use terms reserved for animals.
- No Pronouns:** Do NOT include pronouns (like He/She/It). Output nouns only.
- Output Format:** Provide a single valid JSON object.

Required JSON Structure:

```
{
  "[CODE]": [
    {
      "id": 1,
      "male": "...",
      "female": "...",
      "en_label": "Man-Woman"
    }
    // Continue for all 10 pairs
  ]
}
```

A.7.2 Phase 1: Multilingual Template Adaptation (English to Indic)

To ensure high-fidelity translations and strict adherence to the morphological constraints detailed in Appendix A.2, we decoupled the translation phase from the combinatorial generation phase. We utilized the following representative prompt to adapt the base English JSON templates into the eight target languages.

Role: You are an expert Computational Linguist and native translator for [Target Language].

Task: I have a base JSON file containing English occupational stereotypes, syntactic connectors, and context clauses for a coreference resolution dataset. I need you to translate the values into [Target Language] while strictly adhering to specific grammatical rules to prevent gender leakage.

Input English JSON:

[Insert English JSON containing Occupations, Connectors, and Contexts]

Strict Translation Constraints:

1. **Connector Clauses (Clause 1):** The translated connectors must be strictly gender-neutral. Do not allow the verb in this clause to inflect for the subject's gender.
2. **Morphological Strategy:** *[Insert language-specific hack here, e.g., "Use the ergative case (ne) to block subject-verb agreement" OR "Use plural/honorific verb endings to mask the singular gender"]*.
3. **Context Clauses (Clause 2):** Translate the contexts naturally, leaving the exact placeholder [pronoun] intact. The context must logically explain the occupation's primary duties.
4. **Formatting:** Return ONLY valid JSON with the exact same keys as the input, appending `_[lang_code]` to the translated keys.

A.7.3 Phase 2: Combinatorial Sentence Generation

Following the manual verification of the language-specific JSON files, we employed a deterministic Python script to construct the final evaluation sentences. By utilizing a rule-based algorithm rather than a generative model for the final assembly, we ensured strict mathematical balance, exact string matching for the evaluation framework, and the complete elimination of hallucination risks.

The script operated on the following combinatorial logic:

- **Pairing Strategy:** Occupations were partitioned by their societal gender stereotype. Each occupation was systematically paired with two opposing occupations from the list

(e.g., a male-stereotyped occupation paired with two female-stereotyped occupations, and vice versa) to ensure equal representation.

- **Template Assignment:** To construct the coreference ambiguity, the script applied the "Urgency" syntactic template exclusively for Target-oriented contexts (Label 0) and the "Assistance" template exclusively for Distractor-oriented contexts (Label 1).
- **Pronoun Insertion & Alignment:** For every occupation pair, the script generated two sentence variants: one resolving the context with a male pronoun and the other with a female pronoun.
- **Labeling:** The script automatically assigned the pro-stereotype or anti-stereotype label by checking if the inserted pronoun's gender matched the predefined societal stereotype of the resolution occupation.

A parallel deterministic script was utilized to generate the 160 control sentences following identical pairing and template rules, ensuring the diagnostic baseline was structurally identical to the primary evaluation set.

A.8 Annotator Guideline Sample

To ensure high data quality and cross-lingual consistency, all annotators were provided with standardized, language-specific guidelines. The following sections outline the core instructions provided to the Telugu annotation team, which served as the template for all other target languages.

A.8.1 Wordlist Guidelines

For the translation of the semantic gender pairs used in the WEAT evaluation, annotators were instructed to adhere to the following strict linguistic constraints:

1. **Nominative Singular Form:** Terms must be provided in their pure dictionary form without case suffixes or plural inflections.
2. **Human Context Restriction:** Translations must strictly reference human entities. If a language distinguishes between human and animal gender markers, the human-specific term is mandatory.
3. **Exclusion of Pronouns:** The translations must be standalone nouns (e.g., "Man"); pronouns (e.g., "He") are strictly prohibited.

A.8.2 Main Sentences Guidelines

Annotators evaluated the generated sentence pairs via a tabular interface containing the sentence, the target occupation, and the distractor occupation. The evaluation protocol required a binary Accept/Reject decision based on a two-step validation process:

Step 1: Syntactic Fluency and Naturalness Annotators verified whether the sentence adhered to standard Telugu grammar and phrasing.

- **Criteria:** Does the sentence flow naturally to a native speaker? Are the case markers and postpositions (e.g., *-ni/-nu* vs. *-ki/-ku*) applied correctly to the English loanwords?
- **Rejection Condition:** Sentences exhibiting awkward phrasing, incorrect verb conjugations, or artifacts of machine translation were rejected under the “Grammar” error category.

Step 2: Semantic Coherence and Resolution Logic Annotators verified the logical integrity of the coreference resolution.

- **Criteria:** Does the pronoun in the subordinate clause (e.g., అతను [He] or ఆమె [She]) logically refer to the intended expected resolution based on the contextual action?
- **Rejection Condition:** If the contextual action contradicts the designated target occupation (e.g., a sentence implying a mechanic is changing a saline bottle), the sentence was rejected under the “Logic” error category.

Control Set Specifics For the syntactically resolvable control set (IDs: CTRL_0xxx), annotators were instructed to enforce strict grammatical gender agreement. Because the control set utilizes natively inflected occupational nouns (e.g., గాయకుడు [Male Singer]), the resolving pronoun in the subordinate clause must exactly match the grammatical gender of the antecedent noun. Any mismatch resulted in an immediate rejection.

A.8.3 Recruitment and Disclaimers

Annotators were recruited based on native-level proficiency in their respective target languages and professional backgrounds in linguistics, translation, or data annotation. Prior to beginning the task, all annotators were required to read and digitally sign the following informed consent and content warning disclaimer:

Informed Consent and Study Context

Project Purpose: You are being invited to participate in a research study aimed at evaluating artificial intelligence models for occupational gender bias. Your task is to verify the grammatical correctness and logical resolution of translated sentences.

Study Context: To effectively test these AI models, this dataset contains sentences that pair specific occupations with gendered pronouns. Some of these pairings reflect traditional societal gender stereotypes (e.g., referring to a mechanic as “he” or a receptionist as “she”), while others deliberately counter them. These sentences describe mundane, everyday occupational tasks and are constructed strictly for diagnostic testing. They do not reflect the personal beliefs of the research team.

Privacy & Compensation: Your participation is entirely voluntary, and you may withdraw at any time without penalty. You will be compensated for your time at the agreed-upon professional rate. No personally identifiable information (PII) will be collected, stored, or published in connection with your annotations.

By proceeding to the annotation interface, you acknowledge the context of this study and consent to participate.

A.9 Consistency Metrics

To ensure that observed model performance reflects genuine semantic reasoning rather than heuristic shortcuts, we adapt the consistency metrics introduced in the WINOPRON framework (Gautam et al., 2024). Traditional aggregate metrics, such as the macro-level accuracy gap, can occasionally obscure underlying model incompetence. For instance, a model that indiscriminately selects the first entity in a sentence regardless of context may falsely appear unbiased under standard macro-averages. To decouple true coreference capability from positional or lexical anchoring, we evaluate models on two strict group-level constraints. Pronoun consistency requires a model to correctly resolve the coreference for a specific context across all variations of the gendered pronoun. Complementary to this, disambiguation consistency measures a model’s ability to interpret structural nuance by requiring it to successfully resolve a fixed pronoun to competing antecedents across paired contextual templates. By enforcing success across these paired conditions, these metrics verify whether a system relies on the disambiguating context or simply falls back on rigid occupational stereotypes.

Formally, let x denote a specific contextual template, $p \in \{m, f\}$ represent the pronoun gender (male or female), and $o \in \{t, d\}$ denote the occupational resolution (target or distractor). Let the indi-

cator function $I(x, p, o) \in \{0, 1\}$ denote whether the model correctly resolves a given instance. The Pronoun Consistency (PC) for a specific scenario and occupation is defined as the logical conjunction of correctness across both pronoun genders:

$$\text{PC}(x, o) = I(x, m, o) \wedge I(x, f, o) \quad (1)$$

Similarly, Disambiguation Consistency (DC) evaluates structural reasoning by enforcing correctness across competing antecedents for a fixed pronoun:

$$\text{DC}(x, p) = I(x, p, t) \wedge I(x, p, d) \quad (2)$$

The final reported metrics represent the expected values (percentages) of these strict constraints across the evaluation corpus.

Our empirical observations across 128 evaluations reveal a severe discrepancy between these two metrics. While the mean Pronoun Consistency stands at 43.05%, the mean Disambiguation Consistency plummets to 18.79%. This massive drop-off indicates that models predominantly rely on positional distributions or occupational stereotypes rather than genuine structural reasoning. Furthermore, the analysis exposes stark architectural disparities in coreference competence, as highlighted in Table 10. Models such as Sarvam and Airavata exhibit elevated Pronoun Consistency (50.47% and 46.88%, respectively) but critically fail on strict disambiguation (14.34% and 8.06%), demonstrating an over-reliance on heuristic guessing. Conversely, instruction-tuned models with higher parameter counts, such as Gemma3-27B and Llama3, achieve notably higher Disambiguation Consistency (33.59% and 33.00%, respectively), indicating a much stronger capacity to extract syntactic signals from the disambiguating context.

A.10 Regional Deviation Analysis

To investigate whether occupational stereotypes manifest uniformly across linguistic lines or exhibit localized variations, we introduce a three-step regional deviation analysis. This approach isolates language-specific stereotype reliance from general architectural artifacts and baseline task incompetence. Standard macro-averaging can produce an illusion of neutrality if models lack the syntactic capability to process the prompt, resulting in random guessing that artificially compresses the accuracy gap. To mitigate this, we restrict our analysis strictly to syntactically competent models, defined as those achieving a minimum accuracy of 70% on

the explicitly marked grammatical control set for the target language.

For each target occupation O in language L , we calculate the individual model’s accuracy gap, defined as the pro-stereotypical accuracy minus the anti-stereotypical accuracy. We then aggregate these scores across the K competent models to compute a robust Ensemble Accuracy Gap:

$$\overline{\Delta\text{Acc}}_L(O) = \frac{1}{K} \sum_{m=1}^K \Delta\text{Acc}_{m,L}(O) \quad (3)$$

To quantify how a specific language diverges from the broader multilingual consensus, we subsequently calculate the Regional Mean, $\mu(O)$, by averaging the ensemble gaps across all evaluated languages. The Regional Deviation and its corresponding standard score (Z-Score) are then computed to identify statistically notable localized anchoring:

$$\text{Deviation}_L(O) = \overline{\Delta\text{Acc}}_L(O) - \mu(O) \quad (4)$$

$$\text{Z-Score}_L(O) = \frac{\text{Deviation}_L(O)}{\sigma(O)} \quad (5)$$

Our exploratory analysis utilizing this framework reveals that stereotype reliance is highly localized. Notably, Telugu exhibits a severe average ensemble accuracy gap of +47.94% across all occupations, drastically exceeding the second-highest language, Malayalam, which demonstrates a gap of +8.28%. Furthermore, occupations such as Carpenter, Engineer, and Nurse demonstrate the highest cross-lingual variance, yielding standard deviations of 27.10, 26.42, and 26.15 respectively. These metrics strongly indicate that the strength of certain stereotypes is not a universal artifact of the architecture, but is heavily dictated by the specific sociolinguistic footprint of the target language. Table 11 summarizes the maximum regional deviation observed for each evaluated language.

Table 5: Extrinsic evaluation results...

Table 6: Extrinsic evaluation results across models and languages (Part 1 of 2). Accuracy Gap = Pro Accuracy – Anti Accuracy.

Model	Language	Control	Pro	Anti	Gap
Sarvam-2B	Bengali	0.525	0.515	0.51	0.005
	Hindi	0.5375	0.5275	0.53	-0.0025
	Kannada	0.61875	0.6	0.5625	0.0375
	Malayalam	0.58125	0.5475	0.495	0.0525
	Marathi	0.51875	0.6225	0.5675	0.055
	Odia	0.58125	0.52	0.4875	0.0325
	Tamil	0.59375	0.615	0.6075	0.0075
	Telugu	0.49375	0.62	0.555	0.065
Gemma-2-2B-it	Bengali	0.79375	0.62	0.46	0.16
	Hindi	0.70625	0.4625	0.465	-0.0025
	Kannada	0.83125	0.7125	0.4075	0.305
	Malayalam	0.79375	0.675	0.4275	0.2475
	Marathi	0.6875	0.585	0.505	0.08
	Odia	0.50625	0.495	0.45	0.045
	Tamil	0.7625	0.655	0.5375	0.1175
	Telugu	0.8875	0.7975	0.3575	0.44
Mistral-7B-Instruct	Bengali	0.73125	0.635	0.395	0.24
	Hindi	0.8375	0.5525	0.5375	0.015
	Kannada	0.6875	0.5175	0.45	0.0675
	Malayalam	0.5	0.5175	0.345	0.1725
	Marathi	0.69375	0.54	0.5575	-0.0175
	Odia	0.48125	0.5225	0.495	0.0275
	Tamil	0.55625	0.505	0.51	-0.005
	Telugu	0.50625	0.52	0.58	-0.06
Phi-2	Bengali	0.4625	0.505	0.4975	0.0075
	Hindi	0.4875	0.4875	0.515	-0.0275
	Kannada	0.49375	0.5175	0.51	0.0075
	Malayalam	0.5	0.5	0.4875	0.0125
	Marathi	0.4875	0.5125	0.5425	-0.03
	Odia	0.50625	0.51	0.4975	0.0125
	Tamil	0.49375	0.4675	0.485	-0.0175
	Telugu	0.5	0.4975	0.48	0.0175
Airavata	Bengali	0.5375	0.505	0.4975	0.0075
	Hindi	0.85625	0.6775	0.5325	0.145
	Kannada	0.50625	0.5125	0.5325	-0.02
	Malayalam	0.4875	0.52	0.505	0.015
	Marathi	0.58125	0.5	0.49	0.01
	Odia	0.5125	0.4975	0.505	-0.0075
	Tamil	0.5125	0.4775	0.505	-0.0275
	Telugu	0.55625	0.53	0.5275	0.0025
Aya-23-8B	Bengali	0.5375	0.5325	0.5125	0.02
	Hindi	0.9125	0.7875	0.6725	0.115
	Kannada	0.56875	0.4625	0.485	-0.0225
	Malayalam	0.625	0.5675	0.55	0.0175
	Marathi	0.5875	0.6825	0.6575	0.025
	Odia	0.4875	0.4325	0.4175	0.015
	Tamil	0.65	0.5	0.485	0.015
	Telugu	0.49375	0.3575	0.49	-0.1325
Meta-Llama-3-8B-Instruct	Bengali	0.89375	0.9175	0.5725	0.345
	Hindi	0.89375	0.6825	0.4725	0.21
	Kannada	0.79375	0.8175	0.47	0.3475
	Malayalam	0.85	0.885	0.4875	0.3975
	Marathi	0.74375	0.845	0.5725	0.2725
	Odia	0.79375	0.6675	0.4325	0.235
	Tamil	0.86875	0.8425	0.74	0.1025
	Telugu	0.9875	0.8825	0.2225	0.66
Qwen2.5-7B-Instruct	Bengali	0.6875	0.635	0.2875	0.3475
	Hindi	0.69375	0.515	0.2625	0.2525
	Kannada	0.5	0.4375	0.27	0.1675
	Malayalam	0.70625	0.7525	0.535	0.2175
	Marathi	0.58125	0.5475	0.5075	0.04
	Odia	0.61875	0.5675	0.54	0.0275
	Tamil	0.6375	0.6775	0.5975	0.08
	Telugu	0.64375	0.4925	0.4875	0.005

Table 7: Extrinsic evaluation results across models and languages (Part 2 of 2).

Model	Language	Control	Pro	Anti	Gap
Llama-3.1-8B-Instruct	Bengali	0.875	0.8375	0.4075	0.43
	Hindi	0.60625	0.5475	0.3775	0.17
	Kannada	0.81875	0.9	0.5325	0.3675
	Malayalam	0.8375	0.91	0.6775	0.2325
	Marathi	0.8375	0.8525	0.39	0.4625
	Odia	0.8625	0.745	0.4125	0.3325
	Tamil	0.90625	0.835	0.665	0.17
	Telugu	0.975	0.7725	0.255	0.5175
Tamil-Llama-7B	Bengali	0.56875	0.525	0.49	0.035
	Hindi	0.64375	0.5825	0.57	0.0125
	Kannada	0.4875	0.575	0.5475	0.0275
	Malayalam	0.4625	0.4625	0.445	0.0175
	Marathi	0.525	0.5025	0.485	0.0175
	Odia	0.50625	0.51	0.5025	0.0075
	Tamil	0.61875	0.6325	0.5425	0.09
	Telugu	0.525	0.605	0.5575	0.0475
Yi-1.5-9B-Chat	Bengali	0.7875	0.5425	0.4975	0.045
	Hindi	0.7125	0.5125	0.4625	0.05
	Kannada	0.66875	0.5375	0.4225	0.115
	Malayalam	0.7	0.5175	0.595	-0.0775
	Marathi	0.625	0.53	0.5225	0.0075
	Odia	0.75	0.6175	0.445	0.1725
	Tamil	0.68125	0.61	0.5475	0.0625
	Telugu	0.76875	0.75	0.65	0.1
AryaBhatta-GemmaOrca	Bengali	0.725	0.75	0.365	0.385
	Hindi	0.51875	0.495	0.38	0.115
	Kannada	0.56875	0.78	0.25	0.53
	Malayalam	0.8	0.81	0.4675	0.3425
	Marathi	0.625	0.57	0.6925	-0.1225
	Odia	0.7625	0.79	0.4275	0.3625
	Tamil	0.69375	0.7075	0.3725	0.335
	Telugu	0.85625	0.915	0.3125	0.6025
Sarvam-30B	Bengali	0.5875	0.55	0.4125	0.1375
	Hindi	0.46875	0.46	0.4525	0.0075
	Kannada	0.4625	0.4225	0.4975	-0.075
	Malayalam	0.44375	0.4975	0.4475	0.05
	Marathi	0.4875	0.53	0.5175	0.0125
	Odia	0.53125	0.5875	0.4975	0.09
	Tamil	0.5375	0.45	0.4975	-0.0475
	Telugu	0.44375	0.46	0.435	0.025
Qwen2.5-32B-Instruct	Bengali	0.98125	0.94	0.3775	0.5625
	Hindi	0.975	0.645	0.495	0.15
	Kannada	0.75	0.805	0.385	0.42
	Malayalam	0.95	0.96	0.5625	0.3975
	Marathi	0.825	0.85	0.6675	0.1825
	Odia	0.8625	0.79	0.3025	0.4875
	Tamil	0.8375	0.87	0.5825	0.2875
	Telugu	0.8125	0.8325	0.285	0.5475
Gemma-3-27B-it	Bengali	0.925	0.9425	0.3875	0.555
	Hindi	0.925	0.8875	0.6475	0.24
	Kannada	0.93125	0.985	0.5925	0.3925
	Malayalam	0.99375	0.9725	0.23	0.7425
	Marathi	0.96875	0.9875	0.24	0.7475
	Odia	0.8875	0.8525	0.275	0.5775
	Tamil	0.9875	0.95	0.3575	0.5925
	Telugu	0.9875	1	0.325	0.675
Aya-Expanse-32B	Bengali	0.83125	0.77	0.295	0.475
	Hindi	0.95	0.8225	0.5225	0.3
	Kannada	0.79375	0.76	0.3225	0.4375
	Malayalam	0.9625	0.8775	0.2725	0.605
	Marathi	0.68125	0.6525	0.6575	-0.005
	Odia	0.8125	0.825	0.395	0.43
	Tamil	0.9125	0.88	0.44	0.44
	Telugu	0.74375	0.8725	0.58	0.2925

Table 8: Intrinsic evaluation results across models and languages (Part 1 of 2).

Model	Language	WEAT Effect Size	WEAT p -value	DirectBias
Sarvam-2B	Bengali	0.5781	0.1085	0.0334
	Hindi	0.4788	0.1465	0.0216
	Kannada	0.6909	0.0641	0.0182
	Malayalam	-0.9819	0.9859	0.037
	Marathi	0.0632	0.4465	0.0376
	Odia	0.1312	0.3874	0.0242
	Tamil	0.8223	0.0263	0.0244
	Telugu	0.1942	0.3274	0.0193
Gemma-2-2B-it	Bengali	0.6016	0.0922	0.0431
	Hindi	0.4231	0.1781	0.0604
	Kannada	0.2869	0.2671	0.0596
	Malayalam	-0.3699	0.7948	0.0796
	Marathi	-0.1442	0.6329	0.09
	Odia	-0.235	0.6797	0.0519
	Tamil	0.3464	0.225	0.0613
	Telugu	-0.1935	0.6563	0.0447
Mistral-7B-Instruct	Bengali	0.1071	0.3981	0.1055
	Hindi	0.8516	0.0261	0.0936
	Kannada	0.5957	0.1002	0.2996
	Malayalam	0.2837	0.2706	0.0756
	Marathi	0.7241	0.0506	0.1586
	Odia	-0.2517	0.7143	0.1741
	Tamil	0.2102	0.3144	0.0852
	Telugu	-0.1422	0.6162	0.3714
Phi-2	Bengali	-0.0389	0.5425	0.2321
	Hindi	0.9395	0.019	0.3319
	Kannada	0.7881	0.0413	0.0462
	Malayalam	0.0736	0.4474	0.0609
	Marathi	0.6992	0.0636	0.1122
	Odia	-0.2996	0.7488	0.2261
	Tamil	0.2469	0.3002	0.1835
	Telugu	0.0545	0.4638	0.0321
Airavata	Bengali	-0.5239	0.8766	0.134
	Hindi	0.0483	0.454	0.0097
	Kannada	0.542	0.1239	0.1418
	Malayalam	0.0742	0.4425	0.1404
	Marathi	0.4377	0.1915	0.0151
	Odia	-0.2181	0.6812	0.2545
	Tamil	0.1211	0.3994	0.078
	Telugu	0.2817	0.2713	0.0884
Aya-23-8B	Bengali	-0.3696	0.7823	0.1261
	Hindi	0.6372	0.0856	0.1333
	Kannada	0.4939	0.1444	0.133
	Malayalam	-0.0096	0.509	0.119
	Marathi	0.936	0.0199	0.171
	Odia	-0.2289	0.6944	0.1332
	Tamil	0.1042	0.4191	0.0685
	Telugu	0.1133	0.406	0.0295
Meta-Llama-3-8B-Instruct	Bengali	0.4714	0.1527	0.1375
	Hindi	0.4028	0.1857	0.0398
	Kannada	0.7466	0.0495	0.2034
	Malayalam	0.4016	0.196	0.1548
	Marathi	1.0537	0.0033	0.0663
	Odia	-0.3853	0.7793	0.136
	Tamil	0.3801	0.2033	0.2822
	Telugu	0.0352	0.4757	0.0965
Qwen2.5-7B-Instruct	Bengali	0.5537	0.119	0.1065
	Hindi	0.751	0.0466	0.065
	Kannada	0.6616	0.0698	0.2587
	Malayalam	0.1471	0.3768	0.0591
	Marathi	0.6914	0.0629	0.1288
	Odia	-0.353	0.7744	0.1191
	Tamil	0.1428	0.3819	0.099
	Telugu	0.3301	0.2414	0.1542

Table 9: Intrinsic evaluation results across models and languages (Part 2 of 2).

Model	Language	WEAT Effect Size	WEAT p -value	DirectBias
Llama-3.1-8B-Instruct	Bengali	0.4639	0.1529	0.1407
	Hindi	0.3889	0.1922	0.0447
	Kannada	0.7339	0.0489	0.2202
	Malayalam	0.3777	0.2106	0.1559
	Marathi	1.041	0.0045	0.0657
	Odia	-0.3972	0.7846	0.1361
	Tamil	0.3879	0.1992	0.288
	Telugu	0.1045	0.4235	0.0957
Tamil-Llama-7B	Bengali	-0.4678	0.8484	0.1373
	Hindi	0.8979	0.0184	0.1188
	Kannada	0.6113	0.0921	0.0854
	Malayalam	0.0459	0.4554	0.1889
	Marathi	0.6396	0.0746	0.1499
	Odia	-0.2435	0.7053	0.2002
	Tamil	0.4199	0.178	0.0102
	Telugu	0.2617	0.299	0.0588
Yi-1.5-9B-Chat	Bengali	-0.0408	0.5449	0.1799
	Hindi	0.9912	0.0107	0.088
	Kannada	0.7583	0.0455	0.0912
	Malayalam	0.0883	0.4409	0.0468
	Marathi	0.5479	0.1193	0.0879
	Odia	-0.3213	0.7613	0.2124
	Tamil	0.1907	0.3502	0.2241
	Telugu	0.0451	0.4753	0.0324
AryaBhatta-GemmaOrca	Bengali	0.5674	0.133	0.4647
	Hindi	0.7622	0.0494	0.6711
	Kannada	0.0756	0.4997	0.4791
	Malayalam	-0.1633	0.7128	0.243
	Marathi	-0.2561	0.7314	0.7043
	Odia	0.0895	0.503	0.7483
	Tamil	-0.3572	0.8407	0.2575
	Telugu	-0.3997	0.8513	0.0128
Sarvam-30B	Bengali	0.5851	0.0926	0.0549
	Hindi	0.4381	0.194	0.0177
	Kannada	0.276	0.2666	0.0975
	Malayalam	-0.7956	0.96	0.1056
	Marathi	0.9799	0.0118	0.0604
	Odia	-0.4193	0.8318	0.039
	Tamil	0.3667	0.2179	0.039
	Telugu	-0.3289	0.7658	0.0251
Qwen2.5-32B-Instruct	Bengali	0.5902	0.1008	0.1172
	Hindi	0.8698	0.0271	0.0787
	Kannada	0.755	0.0446	0.2391
	Malayalam	0.2601	0.2907	0.0575
	Marathi	0.3101	0.2588	0.1153
	Odia	-0.2266	0.6837	0.1176
	Tamil	0.2238	0.3133	0.0985
	Telugu	0.3513	0.2216	0.151
Gemma-3-27B-it	Bengali	0.4363	0.1672	0.014
	Hindi	0.2927	0.3287	0.0232
	Kannada	0.2489	0.2868	0.0399
	Malayalam	-0.3004	0.7401	0.056
	Marathi	0.6763	0.0716	0.0205
	Odia	0.3532	0.2289	0.0909
	Tamil	0.5037	0.1307	0.0234
	Telugu	0.0512	0.4614	0.0325
Aya-Expanse-32B	Bengali	-0.5448	0.8783	0.1722
	Hindi	0.6009	0.0924	0.1347
	Kannada	0.4583	0.1596	0.245
	Malayalam	0.0314	0.4784	0.1248
	Marathi	1.0728	0.0065	0.1913
	Odia	-0.224	0.6845	0.1245
	Tamil	0.1252	0.3991	0.0801
	Telugu	0.2407	0.2961	0.0326

Model	Pronoun Consistency	Disambiguation Consistency
<i>Top Performers (Highest Structural Reasoning)</i>		
Gemma3-27B	37.94%	33.59%
Llama3	48.28%	33.00%
Qwen-32B	44.44%	31.66%
<i>Bottom Performers (Highest Heuristic Reliance)</i>		
Airavata	46.88%	8.06%
Mistral	42.81%	7.97%
Phi2	43.47%	6.88%

Table 10: Extremes in Coreference Consistency. The table contrasts the top three and bottom three models based on Disambiguation Consistency. Models at the bottom exhibit high relative Pronoun Consistency but critically fail strict disambiguation, indicating an over-reliance on positional or occupational heuristics rather than syntactic comprehension.

Language	Most Deviant Occ.	Ensemble Gap	Regional Mean	Deviation	Z-Score
Telugu (te)	Beautician	+54.38%	+12.97%	+41.41%	+2.46
Marathi (mr)	Police	−18.75%	+8.35%	−27.10%	−1.48
Odia (od)	Fashion Designer	−16.43%	−1.49%	−14.94%	−1.23
Kannada (kn)	Flight Attendant	−4.17%	+4.86%	−9.03%	−1.08
Tamil (ta)	Police	+25.00%	+8.35%	+16.65%	+0.91
Hindi (hi)	Plumber	−3.89%	+11.07%	−14.96%	−0.90
Bengali (bn)	Security Guard	+4.44%	+14.04%	−9.60%	−0.72
Malayalam (ml)	Beautician	+4.44%	+12.97%	−8.53%	−0.51

Table 11: Maximum Regional Deviation per Language (Excluding Gujarati). The table highlights the single most deviant occupation (highest absolute Z-score) for each evaluated language relative to the cross-lingual regional baseline. Values represent exploratory directional trends isolated from syntactically competent models.