

# A REVIEW OF EMPIRICAL INVESTIGATION OF BIAS TRANSMISSION MECHANISMS IN LARGE-SCALE LANGUAGE MODELS TRAINED ON MULTILINGUAL LOW-RESOURCE DATA STREAMS

Sanjeev Yadav<sup>1</sup>, Mrs. Arifa Khan<sup>2</sup>

<sup>1</sup>Master of Technology, Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

<sup>2</sup>Assistant Professor, Department of Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

\*\*\*

**Abstract** - Large-scale language models (LLMs) have transformed natural language processing through self-supervised pretraining on massive multilingual corpora. However, the expansion of these systems into multilingual and low-resource linguistic settings has intensified concerns regarding bias transmission and amplification. This review synthesizes current empirical evidence on how biases emerge, propagate, and persist in LLMs trained on multilingual low-resource data streams. Drawing from recent studies in computational linguistics, fairness-aware machine learning, and cross-lingual representation learning, the paper categorizes bias transmission mechanisms into three principal dimensions: data-driven bias, model-centric bias, and cross-lingual transfer bias. Particular attention is given to dataset imbalance, tokenization artifacts, representational entanglement across languages, and the hierarchical dominance of high-resource languages during pretraining. The review critically examines existing bias detection metrics, multilingual benchmark limitations, and mitigation strategies, including data rebalancing, embedding debiasing, and fairness-constrained optimization. Despite notable progress, the literature reveals persistent methodological fragmentation and a lack of standardized evaluation protocols tailored to low-resource contexts. The paper concludes by identifying key research gaps and proposing directions for developing linguistically inclusive, culturally sensitive, and empirically grounded fairness frameworks for next-generation multilingual language models.

**Key Words:** Bias transmission, Large-scale language models, Multilingual NLP, Low-resource languages, Fairness-aware machine learning, Cross-lingual transfer.

## 1. INTRODUCTION

### 1.1 Background: Rise of Large-Scale Pretrained Language Models

#### 1.1.1 Evolution of Transformer-Based Architectures

The emergence of Transformer architectures marked a paradigm shift in natural language processing by replacing recurrent and convolutional structures with attention mechanisms capable of modeling long-range dependencies (Vaswani et al., 2017). This architectural innovation enabled large-scale pretraining on massive corpora using self-

supervised objectives. Models such as BERT introduced masked language modeling for contextual representation learning (Devlin et al., 2019), while autoregressive frameworks such as GPT demonstrated the scalability of generative pretraining (Radford et al., 2019; Brown et al., 2020). These developments established the foundation for contemporary large language models (LLMs) characterized by billions of parameters and extensive data exposure.

#### 1.1.2 Emergence of Multilingual Pretraining

To extend performance beyond English-centric systems, multilingual models such as mBERT and XLM-R were trained on corpora spanning dozens to hundreds of languages (Conneau et al., 2020). These models rely on shared subword vocabularies and parameter sharing to facilitate cross-lingual transfer. While multilingual pretraining improves performance in low-resource languages via transfer from high-resource counterparts, it also introduces representational entanglement that may transmit sociocultural biases across linguistic boundaries (Pires, Schlinger and Garrette, 2019).

## 1.2 Conceptual Definitions

### 1.2.1 Bias in Language Models

Bias in LLMs refers to systematic and unfair associations learned from training data that disproportionately favor or disadvantage specific demographic, cultural, or linguistic groups. Empirical studies have demonstrated that pretrained embeddings encode gender, racial, and religious stereotypes reflective of societal distributions present in corpora (Bolukbasi et al., 2016; Caliskan, Bryson and Narayanan, 2017). In multilingual contexts, bias may manifest as linguistic hierarchy, cultural marginalization, or differential performance across languages.

### 1.2.2 Fairness in Multilingual NLP

Fairness in NLP generally concerns equitable treatment across demographic groups and languages in both representation and downstream task performance (Mehrabi et al., 2021). In multilingual settings, fairness extends to parity in accuracy, error distribution, and semantic representation across typologically diverse languages. However, defining fairness operationally remains

challenging due to differing sociolinguistic norms and contextual meanings across cultures.

### 1.2.3 Low-Resource Languages and Multilingual Models

Low-resource languages are characterized by limited digitized corpora, scarce annotated datasets, and minimal computational infrastructure. Multilingual models attempt to mitigate these constraints through shared parameterization and cross-lingual transfer learning (Wu and Dredze, 2019). Although beneficial, this approach often prioritizes high-resource language structures, potentially reinforcing systemic imbalance in representational quality.

## 1.3 Bias in Multilingual Low-Resource Contexts

### 1.3.1 Data Imbalance and Linguistic Dominance

Training corpora for multilingual LLMs are typically dominated by high-resource languages such as English, Mandarin, and Spanish. This disproportionate representation results in uneven parameter allocation and performance disparities (Bender et al., 2021). Consequently, biases embedded in dominant-language data can propagate into low-resource linguistic spaces through shared embeddings and transfer mechanisms.

### 1.3.2 Sociocultural Sensitivity and Ethical Implications

In low-resource contexts, language technologies often serve communities already marginalized in digital ecosystems. Biased outputs may therefore exacerbate exclusion, misrepresentation, or harmful stereotyping. Empirical audits have shown that cross-lingual models may exhibit higher toxicity or stereotyping rates in certain non-English languages due to insufficient contextual grounding (Nozza, Bianchi and Hovy, 2022). Addressing such issues is essential for equitable AI deployment.

## 1.4 Scope and Objectives of the Review

### 1.4.1 Scope

This review systematically synthesizes empirical research on bias transmission mechanisms in large-scale multilingual language models, with particular emphasis on low-resource data streams. It encompasses data-centric, model-centric, and cross-lingual transfer perspectives, integrating findings from computational linguistics, fairness research, and sociotechnical AI studies.

### 1.4.2 Objectives

The primary objectives are threefold:

- To categorize and critically evaluate documented mechanisms through which bias is encoded and transmitted across languages.
- To assess methodological approaches for measuring bias in multilingual low-resource settings.
- To identify unresolved challenges and propose research directions for constructing linguistically inclusive and fairness-aware large language models.

## 2. CONCEPTUAL FOUNDATIONS

### 2.1 Bias in NLP Systems

#### 2.1.1 Types of Bias in NLP

Bias in NLP systems manifests in multiple interrelated forms. Representational bias refers to the unequal or stereotypical portrayal of social groups within textual corpora and learned embeddings. Empirical evidence shows that word embeddings capture gender and racial stereotypes aligned with societal associations (Bolukbasi et al., 2016). Algorithmic bias arises from model training dynamics and optimization processes that amplify existing disparities, even when data distributions appear neutral. Societal bias reflects historical and cultural inequities embedded in source corpora, particularly web-scale data (Bender et al., 2021). Demographic bias is observed when systems yield systematically different performance outcomes across population groups, languages, or dialects, often disadvantaging marginalized communities (Blodgett et al., 2020). In multilingual settings, these categories overlap, especially when high-resource language norms dominate representational space.

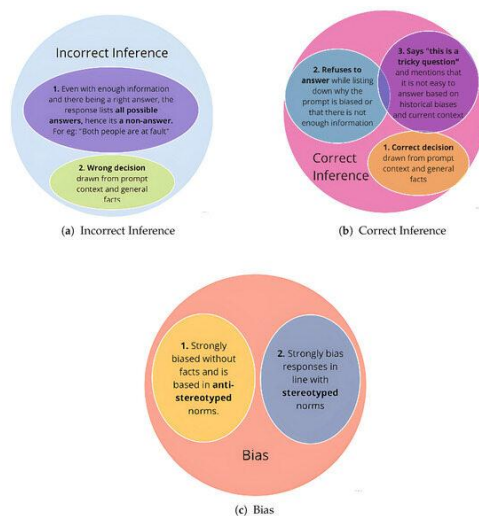


Figure-1: Types of Bias in NLP

### 2.1.2 Metrics for Quantifying Bias

Quantitative bias assessment relies on embedding-based, probabilistic, and task-level metrics. The Word Embedding Association Test (WEAT) measures differential associations between target and attribute word sets (Caliskan, Bryson and Narayanan, 2017). Sentence-level extensions such as SEAT and contextualized embedding probes evaluate bias in transformer-based models. For generative models, bias is often quantified using log-likelihood differentials, toxicity classifiers, and stereotype scoring benchmarks (Nadeem, Bethke and Reddy, 2021). In multilingual contexts, cross-lingual alignment metrics and performance disparity indices are used to evaluate fairness across languages. However, the lack of standardized multilingual benchmarks complicates cross-study comparability.

### 2.1.3 Ethical, Legal, and Social Implications

Bias in NLP systems has significant ethical and regulatory implications. Biased outputs may reinforce harmful stereotypes, propagate misinformation, or marginalize linguistic minorities. From a governance perspective, fairness concerns intersect with emerging AI regulatory frameworks emphasizing accountability, transparency, and non-discrimination. Sociotechnical analyses argue that bias cannot be treated purely as a technical artifact but must be contextualized within broader power structures shaping data production and technological deployment (Birhane and Guest, 2021). In low-resource environments, biased NLP systems may exacerbate digital inequality by misrepresenting under-documented cultures and languages.

## 2.2 Large-Scale Language Models

### 2.2.1 Transformer Architecture

Large-scale language models are primarily built upon the Transformer architecture, which leverages multi-head self-attention mechanisms to capture contextual dependencies across tokens (Vaswani et al., 2017). Unlike recurrent models, Transformers process sequences in parallel, enabling efficient scaling to billions of parameters. Positional encodings preserve sequence order, while stacked encoder-decoder layers enable hierarchical representation learning. The scalability of this architecture has facilitated the training of increasingly large and data-intensive models.

### 2.2.2 Pretraining Paradigms

Two dominant pretraining paradigms underpin modern LLMs. Masked Language Modeling (MLM), introduced in BERT, predicts randomly masked tokens to learn bidirectional contextual representations (Devlin et al., 2019). In contrast, autoregressive next-token prediction, employed in GPT-style models, generates text sequentially by modeling conditional probabilities over tokens (Brown et al., 2020). Variants such as sequence-to-sequence denoising objectives

extend these paradigms for multilingual and generative tasks. While these objectives improve generalization, they may inadvertently encode correlations reflecting biased co-occurrence patterns in large-scale corpora.

### 2.2.3 Multilingual Variants and Cross-Lingual Transfer

Multilingual models such as mBERT and XLM-R are trained on concatenated corpora from multiple languages using shared subword vocabularies (Conneau et al., 2020). Parameter sharing enables cross-lingual transfer, allowing knowledge learned from high-resource languages to benefit low-resource ones. Empirical studies demonstrate emergent cross-lingual alignment without explicit supervision (Pires, Schlinger and Garrette, 2019). However, this shared representation space may also facilitate the transfer of dominant-language biases into underrepresented linguistic contexts, leading to representational imbalance and fairness concerns.

## 2.3 Low-Resource Data Streams

### 2.3.1 Definition of Low-Resource Languages

Low-resource languages are those with limited digitized corpora, scarce annotated datasets, and minimal computational resources for NLP development. These languages often lack standardized orthography, domain-diverse text collections, and large-scale parallel corpora. The disparity in digital presence between high- and low-resource languages creates structural inequities in model performance and representation (Joshi et al., 2020).

### 2.3.2 Challenges of Data Scarcity and Imbalance

Data scarcity introduces both quantitative and qualitative challenges. Limited corpus size restricts vocabulary coverage and contextual diversity, reducing model robustness. Additionally, multilingual training corpora are typically imbalanced, with high-resource languages dominating token distribution. This imbalance leads to disproportionate parameter optimization toward majority languages, resulting in degraded performance and potential bias in minority language outputs (Wu and Dredze, 2019). Noise, code-switching, and orthographic variability further complicate model training in low-resource environments.

### 2.3.3 Synthetic Augmentation and Transfer Learning

To address scarcity, researchers employ synthetic data augmentation methods such as back-translation, machine translation bootstrapping, and data synthesis via generative models. Transfer learning approaches leverage multilingual pretraining to improve downstream performance in low-resource languages. While these techniques enhance accuracy, they may introduce translation artifacts or propagate structural biases from source languages. Recent

studies advocate adaptive fine-tuning, balanced sampling strategies, and culturally grounded data curation to mitigate these risks and improve equitable representation across linguistic communities.

### 3. SURVEY OF BIAS TRANSMISSION MECHANISMS

This section synthesizes empirically identified pathways through which bias is encoded, amplified, and propagated in multilingual large-scale language models trained on low-resource data streams.

#### 3.1 Dataset-Induced Bias

##### 3.1.1 Data Collection Sources

Bias transmission often originates at the data acquisition stage. Large language models are typically pretrained on web-crawled corpora, encyclopedic sources, and user-generated content, which reflect historical and societal asymmetries. Web-scale datasets such as Common Crawl disproportionately represent dominant linguistic and cultural narratives, embedding hegemonic perspectives into training corpora (Bender et al., 2021). For multilingual corpora, disparities in digital presence mean that high-resource languages contribute substantially more textual volume and topical diversity than low-resource languages, predisposing models toward skewed representations.

##### 3.1.2 Sampling Imbalance Across Languages

Token-level imbalance across languages constitutes a primary mechanism of bias transmission. During multilingual pretraining, sampling strategies often allocate training probability proportional to corpus size, favoring high-resource languages. Empirical studies demonstrate that such imbalance leads to uneven parameter updates and performance gaps across languages (Conneau et al., 2020). Temperature-based sampling partially mitigates dominance effects, yet residual disparities persist, particularly for morphologically rich or underrepresented languages (Arivazhagan et al., 2019). Consequently, low-resource languages may inherit structural biases embedded in majority-language distributions.

##### 3.1.3 Cultural and Sociolinguistic Skew

Beyond quantitative imbalance, qualitative cultural skew shapes representational bias. Linguistic corpora frequently overrepresent urban, standardized, or majority dialects while marginalizing regional or minority speech forms. This imbalance results in sociolinguistic underrepresentation and misclassification in downstream tasks (Blodgett et al., 2020). In multilingual contexts, culturally specific meanings may be misaligned or flattened during joint training, leading to semantic homogenization that privileges dominant cultural frameworks.

#### 3.2 Model Training Dynamics

##### 3.2.1 Pre training Objectives and Bias Amplification

Self-supervised objectives such as masked language modeling and autoregressive next-token prediction optimize likelihood-based learning without fairness constraints. These objectives reinforce high-frequency associations present in training data, thereby amplifying stereotypical correlations (Zhao et al., 2019). In multilingual settings, biased co-occurrence patterns in high-resource languages may be statistically reinforced and propagated through shared representations, increasing the persistence of demographic stereotypes.

##### 3.2.2 Cross-Lingual Transfer Effects

Cross-lingual transfer enables performance gains in low-resource languages through shared embedding spaces. However, empirical analyses reveal that representational alignment may transmit biases from dominant languages into structurally distinct ones (Pires, Schlinger and Garrette, 2019). For example, gender associations embedded in English corpora can influence representation in languages with grammatical gender systems, producing compounded bias effects. Such transfer-induced bias is particularly pronounced when low-resource data lacks sufficient counterbalancing examples.

##### 3.2.3 Parameter Sharing and Representation Entanglement

Multilingual LLMs rely on shared parameters across languages to achieve scalability. While efficient, this strategy induces representational entanglement, whereby latent features encode overlapping linguistic and cultural signals. Research indicates that entangled representations can obscure language-specific nuances and propagate majority-language dominance in embedding geometry (Wu and Dredze, 2019). This structural coupling serves as a conduit for bias transmission, particularly when model capacity is insufficient to differentiate diverse linguistic patterns.

#### 3.3 Architecture and Optimization

##### 3.3.1 Attention Mechanisms and Contextual Bias

Attention mechanisms dynamically weight contextual tokens, influencing how associations are formed. Analyses of transformer attention patterns suggest that attention heads may disproportionately focus on socially salient or stereotypically associated tokens (Vig et al., 2020). Such contextual weighting can magnify biased associations during both pretraining and inference. In multilingual models, attention distributions may vary across languages, leading to inconsistent semantic emphasis and uneven bias manifestation.

### 3.3.2 Tokenization Artifacts and Subword Bias

Subword tokenization schemes, such as Byte-Pair Encoding and SentencePiece, create shared vocabularies across languages. While enabling efficient parameter sharing, these methods may fragment low-resource language words into disproportionately many subword units, reducing representational fidelity (Rust et al., 2021). Tokenization artifacts can distort semantic coherence and introduce frequency-based distortions, effectively privileging morphologically simpler or high-frequency language segments.

### 3.3.3 Loss Functions and Optimization Biases

Optimization strategies influence how gradients are allocated across languages and examples. Standard cross-entropy loss functions treat all tokens equally, regardless of language representation or social sensitivity. Without reweighting mechanisms, optimization tends to prioritize frequent patterns, reinforcing majority-language dominance. Adaptive training methods, including fairness-aware regularization, have been proposed to counteract such biases, though empirical validation in multilingual low-resource settings remains limited (Mehrabi et al., 2021).

## 3.4 Evaluation and Benchmarking Bias

### 3.4.1 Core Evaluation Benchmarks

Bias evaluation commonly employs benchmark datasets designed to probe stereotypes, toxicity, or fairness-related disparities. Resources such as StereoSet and CrowS-Pairs measure stereotypical associations in generative and masked language models (Nadeem, Bethke and Reddy, 2021). However, these benchmarks are predominantly English-centric, limiting their generalizability to multilingual contexts. Cross-lingual evaluation frameworks are emerging but remain uneven in linguistic coverage.

### 3.4.2 Limitations of Current Metrics in Multilingual Low-Resource Settings

Existing bias metrics often assume culturally consistent semantic categories, which may not hold across languages. Direct translation of benchmarks can introduce semantic drift and fail to capture culturally specific stereotypes. Furthermore, low-resource languages frequently lack gold-standard annotated fairness datasets, constraining reliable empirical analysis. Scholars argue that multilingual fairness assessment requires culturally grounded evaluation design and community-informed annotation practices to ensure contextual validity (Birhane and Guest, 2021). Without such adaptations, benchmarking processes risk underestimating bias severity in marginalized linguistic settings.

## 4. LITERATURE REVIEW

This section organizes prior scholarship chronologically and thematically to trace the evolution of bias research from monolingual embeddings to multilingual large-scale language models, with emphasis on low-resource contexts.

### 4.1 Early Work on Bias in Monolingual Language Models

#### 4.1.1 Foundational Bias Studies in English Models

Initial empirical investigations into bias in NLP focused primarily on English word embeddings. A seminal contribution was the Word Embedding Association Test (WEAT), which demonstrated that distributional embeddings encode human-like implicit biases, particularly along gender and racial dimensions (Caliskan, Bryson and Narayanan, 2017). Earlier work showed that analogical reasoning in embeddings reproduced stereotypical associations, such as linking professions with specific genders (Bolukbasi et al., 2016). These studies established that statistical co-occurrence patterns in corpora are sufficient to encode socially salient stereotypes, even without explicit labeling.

#### 4.1.2 Early Bias Detection and Mitigation Techniques

Following detection, research turned toward mitigation strategies. Hard and soft debiasing methods were proposed to remove gender subspaces from static embeddings (Bolukbasi et al., 2016). Subsequent work introduced counterfactual data augmentation and adversarial training to reduce demographic signal leakage in contextual models (Zhao et al., 2018). However, later analyses argued that many debiasing methods reduce measurable bias without fully eliminating latent associations, suggesting the persistence of deeper representational issues (Gonen and Goldberg, 2019). These early findings laid the methodological groundwork for bias evaluation in more complex architectures.

### 4.2 Emergence of Multilingual Models

#### 4.2.1 Development of Multilingual Pretrained Architectures

The introduction of multilingual BERT (mBERT) marked a transition toward cross-lingual pretrained representations trained on jointly concatenated corpora (Devlin et al., 2019). Cross-lingual Language Model (XLM) and its robust variant XLM-R further expanded coverage and improved alignment across 100 languages (Conneau et al., 2020). Encoder-decoder architectures such as mT5 extended multilingual modeling into generative and translation tasks (Xue et al., 2021). These models demonstrated emergent cross-lingual transfer without explicit alignment objectives.

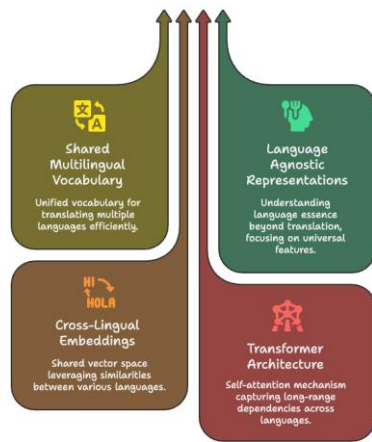


Figure-2: Multilingual Model

#### 4.2.2 Transfer Learning and Language Hierarchy Effects

Empirical studies observed that multilingual models exhibit hierarchical language performance patterns, with high-resource languages benefiting disproportionately from model capacity (Wu and Dredze, 2019). Cross-lingual probing experiments revealed that alignment quality depends on typological similarity and data volume, often privileging Indo-European languages. This “language hierarchy effect” implies that multilingual pretraining may structurally reinforce dominance relationships among languages, raising concerns about fairness in low-resource contexts.

#### 4.3 Bias Analyses in Low-Resource Settings

##### 4.3.1 Studies on African, South Asian, and Indigenous Languages

More recent research has shifted focus toward underrepresented linguistic communities. Analyses of multilingual embeddings have uncovered differential stereotype patterns in African and South Asian languages, often shaped by translation artifacts and sociocultural variation (Nekoto et al., 2020). Investigations into Indigenous language modeling highlight both data scarcity and cultural misrepresentation as sources of bias. These studies emphasize that bias manifestations are context-specific and cannot be assumed to mirror English-language patterns.

##### 4.3.2 Quantitative versus Qualitative Detection Approaches

Quantitative bias detection relies on translated benchmarks and embedding association tests adapted for multilingual

evaluation. However, scholars argue that purely quantitative metrics may overlook culturally embedded meanings and localized forms of discrimination (Blodgett et al., 2020). Qualitative analyses, including community-informed annotation and discourse-level examination, provide complementary insights into nuanced bias expressions. The literature increasingly supports mixed-method approaches to ensure contextual validity in low-resource settings.

#### 4.4 Cross-Lingual and Multilingual Bias Transmission

##### 4.4.1 Comparative Cross-Language Analyses

Comparative evaluations across languages reveal that bias levels vary depending on linguistic structure, corpus composition, and cultural framing. Studies comparing gender bias across multiple European and Asian languages demonstrate both shared and language-specific stereotype patterns (Lauscher and Glavaš, 2019). Such findings indicate that multilingual models do not uniformly distribute bias but instead reshape it according to representational alignment dynamics.

##### 4.4.2 Evidence of Bias Propagation from High- to Low-Resource Languages

Research on cross-lingual transfer suggests that biases present in dominant-language corpora can propagate into low-resource languages via shared embeddings and parameter coupling (Pires, Schlinger and Garrette, 2019). For instance, stereotypical occupational associations encoded in English may influence predictions in typologically distinct languages lacking equivalent data distributions. This propagation effect underscores the role of multilingual joint training as a structural mechanism for bias transmission rather than merely a passive reflection of local corpora.

#### 4.5 Bias Mitigation Techniques

##### 4.5.1 Data-Centric Approaches

Data-centric mitigation strategies aim to address bias at the source. Techniques include corpus balancing, oversampling underrepresented groups, and curating culturally diverse datasets. Temperature-based resampling has been employed to reduce dominance effects in multilingual pretraining (Arivazhagan et al., 2019). Counterfactual data augmentation further attempts to neutralize gendered or demographic associations at scale. However, ensuring cultural authenticity while balancing representation remains an open challenge.

##### 4.5.2 Model-Centric Methods

Model-centric interventions operate within training or fine-tuning processes. Approaches include adversarial debiasing, fairness-aware regularization, and constrained optimization

techniques designed to minimize demographic signal retention (Mehrabi et al., 2021). Recent multilingual research explores disentangled representation learning to separate language identity from sociocultural attributes. Despite promising results, empirical evidence suggests that mitigation often reduces surface-level bias metrics while deeper structural associations persist, indicating the need for more theoretically grounded fairness frameworks.

## 5. THEMATIC SYNTHESIS

This section integrates the reviewed literature to identify convergent patterns, methodological tensions, and structural insights regarding bias transmission in multilingual large-scale language models trained on low-resource data streams.

### 5.1 Emerging Patterns Across Empirical Studies

#### 5.1.1 Systematic Dominance of High-Resource Languages

A consistent pattern across empirical investigations is the structural dominance of high-resource languages in multilingual pretraining. Studies demonstrate that token imbalance and corpus heterogeneity lead to disproportionate parameter allocation, yielding higher representational fidelity and downstream performance for majority languages (Conneau et al., 2020). This dominance effect extends beyond accuracy disparities, influencing embedding geometry and semantic alignment. As a result, biases embedded in high-resource corpora often shape the shared latent space of multilingual models.

#### 5.1.2 Amplification of Societal Stereotypes

Research further indicates that large-scale pretraining amplifies pre-existing societal stereotypes rather than merely reflecting them. Likelihood-based optimization reinforces high-frequency co-occurrences, intensifying stereotypical associations related to gender, race, or profession (Zhao et al., 2019). In multilingual contexts, these amplified associations may permeate linguistically distinct environments through shared parameters, producing translingual bias patterns even where local corpora lack equivalent distributions.

### 5.2 Conflicting Results and Methodological Divergence

#### 5.2.1 Variability in Bias Measurement Outcomes

The literature reports inconsistent bias magnitudes depending on evaluation methodology. Embedding association tests often detect significant stereotype effects, while task-based evaluations sometimes reveal weaker or context-dependent disparities (Gonen and Goldberg, 2019). Differences in template design, translation fidelity, and annotation quality contribute to divergent findings. In

multilingual low-resource settings, translated benchmarks may distort semantic nuance, leading to measurement artifacts rather than genuine bias estimation.

#### 5.2.2 Disagreement on Transfer Effects

Scholarly disagreement also exists regarding the extent to which cross-lingual transfer propagates bias. Some studies argue that multilingual alignment facilitates bias diffusion from dominant languages (Pires, Schlinger and Garrette, 2019), whereas others report language-specific attenuation effects depending on morphological or syntactic divergence. Such discrepancies often arise from differences in corpus composition, sampling temperature, and model capacity, indicating the absence of standardized experimental protocols for multilingual fairness research.

### 5.3 Strengths and Weaknesses of Current Approaches

#### 5.3.1 Strengths: Scalability and Diagnostic Innovation

Current research demonstrates methodological sophistication in bias diagnostics. The development of benchmark datasets such as StereoSet and multilingual probing tasks has enabled systematic evaluation across architectures (Nadeem, Bethke and Reddy, 2021). Moreover, multilingual pretraining offers scalability advantages, enabling low-resource languages to benefit from shared representations. Data augmentation and adaptive sampling strategies further illustrate innovative attempts to address imbalance at scale.

#### 5.3.2 Weaknesses: Benchmark Limitations and Context Insensitivity

Despite progress, existing approaches exhibit notable limitations. Many fairness benchmarks are English-centric, and direct translation often fails to capture culturally specific stereotypes or linguistic subtleties (Blodgett et al., 2020). Additionally, mitigation techniques frequently target surface-level statistical parity without addressing deeper sociotechnical roots of bias. The absence of culturally grounded evaluation frameworks limits interpretability, particularly in Indigenous and under-documented language contexts.

### 5.4 Interplay between Linguistic Typology and Bias Transmission

#### 5.4.1 Typological Features and Representational Alignment

Linguistic typology influences how bias manifests and propagates. Languages differ in morphological complexity, grammatical gender systems, and word order structures. Research indicates that typological similarity facilitates cross-lingual alignment, thereby increasing the likelihood of

bias transfer between structurally related languages (Wu and Dredze, 2019). Conversely, typologically distant languages may experience attenuated but not eliminated bias effects due to shared subword vocabularies and parameter entanglement.

#### 5.4.2 Grammatical Gender and Structural Reinforcement

Languages with grammatical gender systems present unique bias dynamics. When pretrained on mixed corpora, models may conflate grammatical and social gender cues, reinforcing stereotypical occupational or role-based associations (Lauscher and Glavaš, 2019). In contrast, gender-neutral languages may exhibit bias primarily through semantic associations rather than morphological marking. This typological interplay underscores that bias transmission is not solely data-driven but also mediated by structural linguistic properties.

### 6. CONCLUSION

This review synthesized empirical and conceptual research on bias transmission mechanisms in large-scale language models trained on multilingual low-resource data streams. The analysis demonstrates that bias is not confined to isolated model components but emerges from the interaction between data imbalance, pretraining objectives, architectural design, and cross-lingual parameter sharing. High-resource languages exert structural influence during multilingual training, shaping embedding geometry and facilitating the propagation of societal stereotypes into underrepresented linguistic contexts. While significant methodological advancements have been made in bias detection and mitigation, current evaluation frameworks remain predominantly English-centric and insufficiently adapted to culturally diverse environments. Evidence further indicates that typological characteristics, tokenization strategies, and optimization dynamics modulate how bias manifests across languages. Despite progress in fairness-aware modeling and data-centric interventions, mitigation efforts often address surface-level statistical disparities rather than deeper sociotechnical determinants. Overall, the literature underscores the need for standardized multilingual benchmarks, culturally grounded evaluation protocols, and typology-aware fairness frameworks. Advancing equitable multilingual NLP requires interdisciplinary collaboration integrating computational rigor, linguistic insight, and ethical accountability to ensure inclusive and socially responsible language technologies.

### 7. LIMITATIONS OF THE REVIEW

This review is limited by the rapidly evolving nature of large-scale language model research, where new architectures and evaluation benchmarks emerge frequently. Although efforts were made to include diverse multilingual studies, the available literature remains skewed toward widely studied

languages, potentially restricting coverage of extremely low-resource or Indigenous contexts. Additionally, many existing studies rely on heterogeneous experimental settings, limiting direct comparability across findings. The review synthesizes published empirical evidence but does not include meta-analytic statistical aggregation of results. Finally, access to proprietary training data and large-scale model parameters constrains transparency in several referenced works, affecting comprehensive assessment of bias transmission mechanisms.

### References

1. Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M., Cao, Y., Foster, G., Cherry, C. and Macherey, W. (2019) 'Massively multilingual neural machine translation in the wild: Findings and challenges', arXiv preprint arXiv:1907.05019.
2. Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) 'On the dangers of stochastic parrots: Can language models be too big?', Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), pp. 610–623.
3. Birhane, A. and Guest, O. (2021) 'Towards decolonising computational sciences', *Patterns*, 2(10), 100289.
4. Blodgett, S.L., Barocas, S., Daumé III, H. and Wallach, H. (2020) 'Language (technology) is power: A critical survey of "bias" in NLP', Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5454–5476.
5. Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V. and Kalai, A.T. (2016) 'Man is to computer programmer as woman is to homemaker? Debiasing word embeddings', *Advances in Neural Information Processing Systems*, 29, pp. 4349–4357.
6. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020) 'Language models are few-shot learners', *Advances in Neural Information Processing Systems*, 33, pp. 1877–1901.
7. Caliskan, A., Bryson, J.J. and Narayanan, A. (2017) 'Semantics derived automatically from language corpora contain human-like biases', *Science*, 356(6334), pp. 183–186.
8. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V. (2020) 'Unsupervised cross-lingual representation learning at scale', Proceedings of the

- 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451.
9. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', Proceedings of NAACL-HLT 2019, pp. 4171–4186.
  10. Gonen, H. and Goldberg, Y. (2019) 'Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them', Proceedings of NAACL-HLT 2019, pp. 609–614.
  11. Joshi, P., Santy, S., Budhiraja, A., Bali, K. and Choudhury, M. (2020) 'The state and fate of linguistic diversity and inclusion in the NLP world', Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6282–6293.
  12. Lauscher, A. and Glavaš, G. (2019) 'Are we consistently biased? Multidimensional analysis of biases in distributional word vectors', Proceedings of the 2019 Workshop on Gender Bias in Natural Language Processing, pp. 85–91.
  13. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021) 'A survey on bias and fairness in machine learning', ACM Computing Surveys, 54(6), pp. 1–35.
  14. Nadeem, M., Bethke, A. and Reddy, S. (2021) 'StereoSet: Measuring stereotypical bias in pretrained language models', Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, pp. 5356–5371.
  15. Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohunge, T., Akinola, S.O., Muhammad, S., Kabongo, S., Osei, S., Sackey, F. et al. (2020) 'Participatory research for low-resourced machine translation: A case study in African languages', Findings of the Association for Computational Linguistics (EMNLP 2020), pp. 2144–2160.
  16. Pires, T., Schlinger, E. and Garrette, D. (2019) 'How multilingual is multilingual BERT?', Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4996–5001.
  17. Rust, P., Pfeiffer, J., Vulić, I., Ruder, S. and Gurevych, I. (2021) 'How good is your tokenizer? On the monolingual performance of multilingual language models', Proceedings of ACL 2021, pp. 3118–3135.
  18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', Advances in Neural Information Processing Systems, 30, pp. 5998–6008.
  19. Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y. and Shieber, S. (2020) 'Investigating gender bias in language models using causal mediation analysis', Advances in Neural Information Processing Systems, 33.
  20. Wu, S. and Dredze, M. (2019) 'Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT', Proceedings of EMNLP-IJCNLP 2019, pp. 833–844.
  21. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A. and Raffel, C. (2021) 'mT5: A massively multilingual pre-trained text-to-text transformer', Proceedings of NAACL-HLT 2021, pp. 483–498.
  22. Zhao, J., Wang, T., Yatskar, M., Ordonez, V. and Chang, K.-W. (2018) 'Gender bias in coreference resolution: Evaluation and debiasing methods', Proceedings of NAACL-HLT 2018, pp. 15–20.
  23. Zhao, J., Zhou, Y., Li, Z., Wang, W. and Chang, K.-W. (2019) 'Learning gender-neutral word embeddings', Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4847–4853.
  24. Adebayo, J., Muelly, M., Liccardi, I. and Kim, B. (2020) 'Debugging tests for model explanations', Advances in Neural Information Processing Systems, 33, pp. 700–712.
  25. Barocas, S., Hardt, M. and Narayanan, A. (2019) Fairness and Machine Learning: Limitations and Opportunities. Cambridge, MA: fairmlbook.org.
  26. Blodgett, S.L. and O'Connor, B. (2017) 'Racial disparity in natural language processing: A case study of social media African-American English', Proceedings of Fairness, Accountability, and Transparency in Machine Learning (FAT/ML).
  27. Costa-jussà, M.R., Cross, J., Çelebi, O. and Heafield, K. (2020) 'Google's multilingual neural machine translation system: Enabling zero-shot translation', Transactions of the Association for Computational Linguistics, 8, pp. 339–351.
  28. Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D. and Weston, J. (2020) 'Queens are powerful too: Mitigating gender bias in dialogue generation', Proceedings of EMNLP 2020, pp. 8173–8188.
  29. Dodge, J., Gururangan, S., Card, D., Schwartz, R. and Smith, N.A. (2021) 'Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus', Proceedings of EMNLP 2021, pp. 1286–1305.

30. Ethayarajh, K. (2019) 'How contextual are contextualized word representations?', Proceedings of ACL 2019, pp. 55–65.
31. Friedman, B. and Nissenbaum, H. (1996) 'Bias in computer systems', ACM Transactions on Information Systems, 14(3), pp. 330–347.
32. Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J. (2018) 'Word embeddings quantify 100 years of gender and ethnic stereotypes', Proceedings of the National Academy of Sciences, 115(16), pp. E3635–E3644.
33. Hardt, M., Price, E. and Srebro, N. (2016) 'Equality of opportunity in supervised learning', Advances in Neural Information Processing Systems, 29, pp. 3315–3323.
34. Hovy, D. and Spruit, S.L. (2016) 'The social impact of natural language processing', Proceedings of ACL 2016, pp. 591–598.
35. Huang, P.-S., Stoyanov, V. and Zettlemoyer, L. (2019) 'Addressing the data scarcity issue in multilingual language modeling', Proceedings of CoNLL 2019, pp. 1–10.
36. Jiang, Z., Araki, J., Ding, H. and Neubig, G. (2020) 'How can we know what language models know?', Transactions of the Association for Computational Linguistics, 8, pp. 423–438.
37. Liang, P.P., Wu, C., Morency, L.-P. and Salakhutdinov, R. (2020) 'Towards understanding and mitigating social biases in language models', Proceedings of ICML 2020 Workshop on Responsible AI.
38. Nozza, D., Bianchi, F. and Hovy, D. (2021) 'Honest: Measuring hurtful sentence completion in language models', Proceedings of NAACL 2021, pp. 2398–2406.
39. Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020) 'Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing', Proceedings of FAcT 2020, pp. 33–44.
40. Sheng, E., Chang, K.-W., Natarajan, P. and Peng, N. (2019) 'The woman worked as a babysitter: On biases in language generation', Proceedings of EMNLP-IJCNLP 2019, pp. 3407–3412.
41. Strubell, E., Ganesh, A. and McCallum, A. (2019) 'Energy and policy considerations for deep learning in NLP', Proceedings of ACL 2019, pp. 3645–3650.
42. Talat, Z., Rogers, A., Schmaltz, A., Suresh, H., Blodgett, S., Daumé III, H., Choi, Y. and Smith, N.A. (2022) 'A holistic approach to documenting datasets and models for responsible NLP', Communications of the ACM, 65(3), pp. 90–99.
43. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E. and Azhar, F. (2023) 'LLaMA: Open and efficient foundation language models', arXiv preprint arXiv:2302.13971.
44. Ziems, N., Held, W., Shaikh, O., Chen, J., Zhang, D. and Yang, D. (2022) 'Can large language models transform computational social science?', Computational Linguistics, 48(3), pp. 1–29.