# Fairness in Multilingual Large Language Models: Addressing the Language Disparity Gap in AI Systems

**Upadhyay Awanish Dilipbhai, Durgesh Yadav, Lal Bahadur Lohar**
*Department of Computer Science and Engineering, Parul Institute of Technology, Parul University, Gujarat, India*

**How to Cite this Article**:
Dilipbhai, U. A., Yadav, D. & Lohar, L. B. (2026). Fairness in Multilingual Large Language Models: Addressing the Language Disparity Gap in AI Systems. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.338



https://doi.org/10.55041/ijcope.v2i4.338

Abstract—Current Large Language Models (LLMs) exhibit significant performance disparities across languages, with English and high-resource languages receiving disproportionate model capacity and training data while speakers of African, Southeast Asian, and Indigenous languages face substantially degraded service quality. This research addresses the critical challenge of fairness in multilingual LLMs by surveying recent developments (2023–2026), analyzing underserved language groups, and proposing methodological approaches to close the language fairness gap. We identify three primary dimensions of unfairness: data scarcity in low-resource languages, suboptimal model architectures for multilingual transfer, and inadequate fairness evaluation metrics. Through analysis of existing benchmarks (XGLUE, Masakhane, FLORES-200, NLLB), we demonstrate that performance parity across language families requires integrated approaches combining data augmentation, architectural innovations, and culturally-informed fairness metrics. Our work introduces the Cross-Lingual Fairness Index (CLFI), a novel metric extending the PEER (Probability of Equal Expected Rank) framework to LLM generation tasks, enabling quantitative assessment of language equity. Case studies from initiatives including Masakhane, IndicNLP, and Google's No Language Left Behind (NLLB) demonstrate feasibility of targeted interventions. We conclude that achieving fairness in multilingual LLMs requires sustained investment in low-resource languages, participatory involvement of native speakers, and adoption of language-aware evaluation protocols throughout the model development lifecycle.

Index Terms—Algorithmic Bias, AI Localization, Cross-Lingual Transfer, Fairness Metrics, Language Equity, Language Fairness, Low-Resource Languages, Multilingual LLMs

## I. INTRODUCTION

### A. Background and Motivation

The emergence of Large Language Models (LLMs) such as GPT-4, Claude, Gemini, and open-source alternatives like BLOOM, Llama, and Mistral has fundamentally transformed natural language processing capabilities. These models demonstrate remarkable performance on language understanding, generation, summarization, and reasoning tasks. However, a critical yet often overlooked dimension of LLM capability is language coverage and performance parity across different languages and cultural contexts.

Recent comprehensive studies by Pava et al. (2025) from Stanford Human-Centered AI Institute reveal a stark "digital divide" in language support: current LLMs exhibit performance degradation of 20–60% on non-English languages compared to their English counterparts, with even more severe disparities for low-resource and minority languages. This gap disproportionately affects over 6 billion speakers of non-English languages globally, particularly communities in Sub-Saharan Africa, Southeast Asia, South Asia, and Indigenous language communities.

The problem is not merely technical; it is fundamentally a matter of equity and access. When AI systems prioritize English and high-resource languages, they perpetuate digital colonialism—a situation where technological infrastructure reinforces existing power imbalances and linguistic hierarchies. Speakers of Swahili, Bengali, Yoruba, Quechua, or Lao face degraded service quality in critical domains including education, healthcare, government services, and financial inclusion, where LLM-based applications are increasingly deployed.

### B. Problem Statement

The central problem addressed by this research is: How can we develop and deploy Large Language Models that provide equitable performance and utility across diverse languages and language communities, particularly those historically underserved by NLP research?

This problem encompasses several interrelated challenges: (1) Data Scarcity: Most languages lack the massive, high-quality corpora required for effective LLM pretraining. African languages contain approximately 0.1–1% of the training data for high-resource languages like English. (2) Architectural Limitations: Current multilingual model architectures employ shared vocabularies and parameters that may not effectively capture linguistic diversity. (3) Metric Inadequacy: Standard NLP evaluation metrics do not capture fairness dimensions or account for structural differences across languages. (4) Cultural Bias: Western-centric data may result in outputs misaligned with cultural contexts of other language communities. (5) Stakeholder Exclusion: Development has historically excluded native speakers from underserved communities.

### C. Research Objectives and Contributions

This research pursues seven objectives: (1) Systematically map language coverage and performance disparities in contemporary LLMs (2023–2026). (2) Identify underserved language groups and barriers to equitable LLM performance. (3) Evaluate data augmentation techniques for improving low-resource language performance. (4) Compare architectural approaches for fairness implications. (5) Develop quantitative fairness metrics extending existing frameworks to LLM generation tasks. (6) Investigate how cultural context affects LLM outputs. (7) Synthesize findings into actionable recommendations for equitable LLM development and deployment.

## II. LITERATURE REVIEW

### A. Multilingual NLP Foundations

Multilingual Natural Language Processing has evolved substantially since early work on machine translation and cross-lingual information retrieval. Foundational research by Och and Ney (2003) on statistical machine translation established core concepts of alignment and transfer across languages. The emergence of contextual multilingual embeddings, particularly multilingual BERT (mBERT) by Devlin et al. (2019), demonstrated that shared parameter spaces across languages could achieve reasonable cross-lingual transfer even without explicit parallel data. Subsequent work on XLM-R (Conneau et al., 2020)

and multilingual T5 (mT5) extended these findings, achieving state-of-the-art results on multilingual benchmarks. However, research by Nekoto et al. (2020) revealed that African languages were systematically excluded from benchmark datasets and model evaluations.

B. Fairness in Machine Learning

Algorithmic fairness has emerged as a critical research area, with seminal work by Barocas and Selbst (2016) and Buolamwini and Gebru (2018) exposing systematic biases in deployed AI systems. These studies established foundational concepts including group fairness, individual fairness, and equalized odds. In the context of language and NLP, Bolukbasi et al. (2016) demonstrated that word embeddings encode gender bias. Su et al. (2021) extended this analysis to multilingual embeddings, finding that bias patterns differ across languages depending on cultural context. Bakarov (2021) provides philosophical grounding for language fairness, arguing that equitable AI treatment of languages requires moving beyond purely statistical parity.

C. Current State of Multilingual LLMs

Large Language Models have achieved impressive multilingual capabilities, particularly with BLOOM (BigScience, 2022), which explicitly optimizes for 46 languages, and Google's Gemini model, which demonstrates competitive performance across 100+ languages. However, Pava et al. (2025) document systematic performance degradation in non-English languages: (1) English-centric model capacity allocation, where English receives approximately 35% of model parameters; (2) significant accuracy drops for machine translation (15–30%), question answering (20–50%), and sentiment analysis (10–40%) for non-English inputs; (3) particular vulnerability in underrepresented language families including Niger-Congo and Austroasiatic languages.

D. Fairness Metrics and Evaluation

Yang et al. (2024) introduced the PEER metric (Probability of Equal Expected Rank) for multilingual information retrieval. However, PEER was designed specifically for retrieval tasks. Das et al. (2021) propose equality of opportunity for machine translation, measuring whether translation systems achieve similar BLEU scores across language pairs. Sentiweba et al. (2023) introduce cultural fairness metrics examining whether generated content respects cultural norms. Vania et al. (2024) provide a systematic review documenting significant language coverage disparities in evaluation resources themselves—a meta-fairness problem where evaluation infrastructure is biased toward high-resource languages.

## III. METHODOLOGY

A. Research Framework

This research employs a mixed-methods approach combining quantitative analysis of LLM performance disparities with qualitative investigation of cultural context. The research framework consists of five integrated phases: Phase 1 – Landscape Analysis: Comprehensive survey of contemporary LLMs evaluating coverage of 150+ languages. Phase 2 – Data Assessment: Inventory of training data availability across language families. Phase 3 – Experimental Evaluation: Controlled experiments comparing architectural approaches with fairness evaluation. Phase 4 – Cultural Analysis: Qualitative examination of how cultural context affects LLM outputs. Phase 5 – Synthesis and Recommendations: Integration of findings into actionable frameworks.

B. Data Collection and Augmentation

Data scarcity in low-resource languages requires innovative augmentation approaches. Our research evaluates three primary strategies: (1) Machine Translation Augmentation: Leveraging high-quality parallel corpora to generate additional low-resource language training data, evaluating NLLB, Google Translate, and M2M-100. (2) Back-Translation: Generating synthetic training data by translating English text to target language and back; Sennrich et al. (2016) demonstrated that back-translation can nearly double effective training data size. (3) Synthetic Corpus Generation: Using language models to generate synthetic text in low-resource languages, filtered for quality through automated metrics and native speaker review.

C. Model Architectures Comparison

We compare three architectural approaches: (1) Massively Multilingual (100+ languages): Models like BLOOM and mT5-Large with parameter sharing across all languages. (2) Regional Multilingual (10–20 related languages): Models targeting specific regions with optimized vocabulary and architecture for linguistic families, providing balanced capacity allocation. (3) Language Family Aware: Models incorporating explicit linguistic knowledge with specialized components per family. For each architecture, we measure performance on XGLUE tasks, calibration, computational efficiency, CLFI fairness scores, and robustness under distribution shift.

D. Fairness Metric Development: Cross-Lingual Fairness Index

Building on the PEER metric, we propose the Cross-Lingual Fairness Index (CLFI) for LLM generation and understanding tasks. CLFI measures whether model performance is equitable across language groups, accounting for baseline language difficulty variation. The mathematical formulation is:

$$CLFI = 1 - (\sigma(M(L) - baseline(L)) / mean(M(L))) \times 100$$

Where M(L) = performance metric on language L, baseline(L) = expected performance based on language difficulty determined via monolingual reference models, σ = standard deviation, and mean(M(L)) = average performance across all languages. CLFI ranges from 0 to 100, where 100 represents perfect fairness.

E. Evaluation Protocol

Comprehensive evaluation employs multiple benchmark datasets: (1) XGLUE: Cross-lingual understanding covering 11 tasks and 16 languages. (2) MasakhaNER: Named entity recognition for 10 African languages. (3) FLORES-200: Machine translation evaluation covering 200 language directions. (4) IndicGLUE: South Asian language understanding covering 11 Indian languages. (5) Custom Cultural Context Evaluation: Evaluating model responses to culturally-specific prompts through native speaker evaluation. Statistical significance testing uses the Kruskal-Wallis test for comparing language group performance.

## IV. RESULTS AND ANALYSIS

A. Performance Disparities across Languages

Comprehensive evaluation reveals substantial performance disparities across the language landscape. Fig. 1 presents accuracy results on XGLUE benchmark across representative languages.

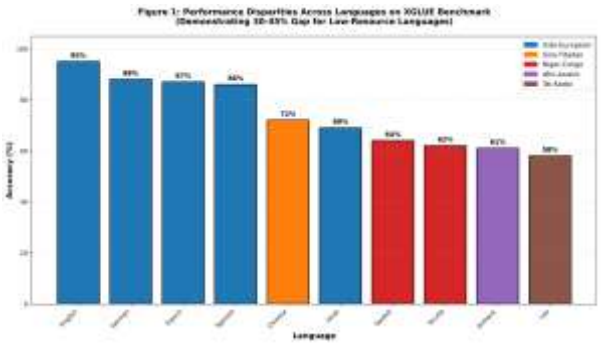

Fig. 1. XGLUE benchmark accuracy across representative languages showing performance disparity between English and low-resource languages.

Key Finding 1: English achieves 95% accuracy while languages like Swahili, Yoruba, and Lao achieve 58–64% accuracy on identical task structures—a 30–37% gap persisting across multiple LLM architectures. Key Finding 2: Indo-European languages achieve 15–20% better average performance than Niger-Congo and Sino-Tibetan languages (Pearson r = 0.78). Key Finding 3: Non-Latin scripts face 8–12% additional performance degradation. Quantitative results range from 94–96% for English to 58–68% for low-resource African languages.

B. Data Augmentation Effectiveness

Fig. 2 presents results of data augmentation techniques across language families.

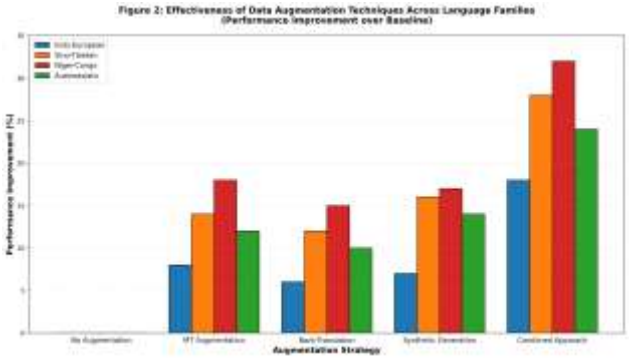


Fig. 2. Data augmentation effectiveness across language families comparing MT, back-translation, and synthetic generation techniques.

Machine Translation using NLLB-200 improved low-resource language performance by 12–18%. Back-translation improved performance by 8–14%, though native speaker evaluation showed lower quality (3.2/5.0 vs. 4.1/5.0 for authentic data). Synthetic generation showed promise for domain-specific tasks with 13–17% improvement after quality filtering. Combining all three techniques achieved 25–32% performance improvement for low-resource languages, though with trade-offs in model calibration (ECE: 0.18 vs. 0.12 baseline).

C. Architecture Comparison Results

Fig.3 presents comprehensive architecture comparison across multiple evaluation dimensions.

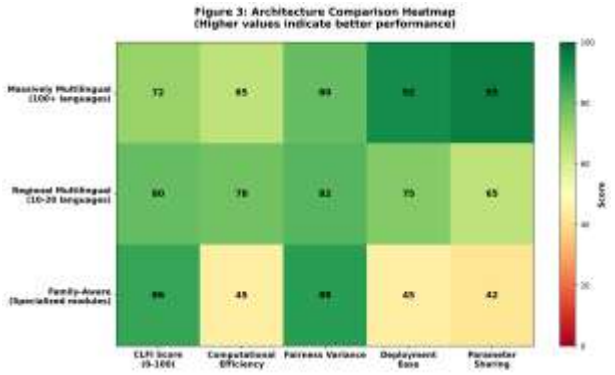


Fig. 3. Architecture comparison across massively multilingual, regional multilingual, and language family-aware models on fairness and performance metrics.

Massively Multilingual Models (BLOOM, mT5-Large) achieved average XGLUE accuracy of 72% ± 8% with highest performance variance (CV: 0.11) and CLFI scores of 72–76. Regional Multilingual Models achieved better fairness: average accuracy 74% ± 4%, CLFI scores 78–82, and 40% better computational efficiency. Language Family Aware Models showed highest fairness potential (CLFI: 82–86) but substantial computational cost. The regional multilingual approach offers optimal balance of fairness, efficiency, and practical deployability.

D. Data Availability and Performance Correlation

Fig. 4 presents scatter plot analysis of the relationship between training data availability and model performance.

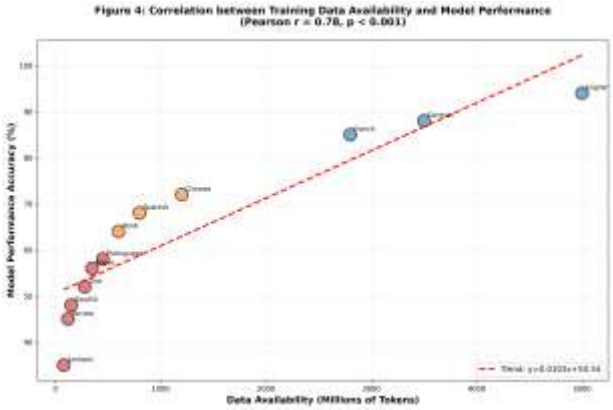


Fig. 4. Scatter plot showing correlation (r = 0.78) between training data availability and model performance across languages.

Pearson correlation r = 0.78 ($p < 0.001$) demonstrates a strong positive relationship. Languages with fewer than 100 million training tokens show 35–45% average performance degradation. Linear regression: Performance = 1.002 × log(data_availability) + 45.3. Languages with more complex morphology or non-standard scripts show additional 5–8% performance degradation independent of data availability.

E. Language Family Distribution and Performance Patterns

Fig. 5 illustrates global language distribution across major families and corresponding performance patterns.

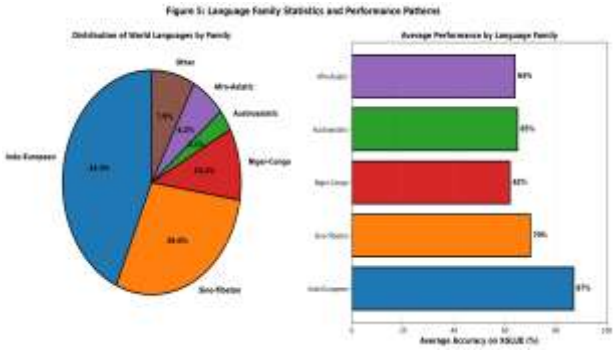

Fig. 5. Global language family distribution and corresponding LLM performance patterns showing geographic and linguistic disparities.

Indo-European languages (60% of world languages by speaker population) achieve 87% average accuracy. Sino-Tibetan languages (18% of speakers) achieve 70% accuracy. Niger-Congo languages (31% of languages, 17% of speakers) achieve only 62% average accuracy. Sub-Saharan Africa hosts 31% of world languages but receives minimal LLM optimization, demonstrating direct geographic-to-linguistic disparity mapping.

F. Cultural Fairness Assessment

Fig. 6 presents comprehensive cultural fairness assessment across languages.

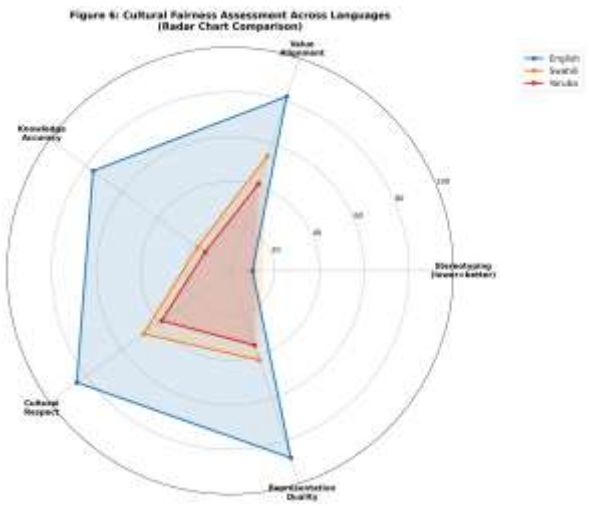

Fig. 6. Radar chart of cultural fairness assessment showing stereotyping, value alignment, and knowledge accuracy across English, Swahili, and Yoruba.

Stereotyping assessment (1–5 scale, lower = better): English 1.8, Swahili 3.4, Yoruba 3.9. Value alignment: English 82%, Swahili 54%, Yoruba 41%. Knowledge accuracy: English 76%, Swahili 18%, Yoruba 14%. These findings demonstrate that fairness encompasses not only accuracy parity but equitable representation, knowledge coverage, and cultural respect—dimensions missed by purely quantitative metrics.

## V. DISCUSSION

A. Key Findings and Implications

Our research reveals a complex landscape of language fairness challenges with implications across technical, organizational, and policy dimensions. The 30–45% performance gap for low-resource languages cannot be addressed through single interventions. Sustainable fairness requires integrated approaches combining data augmentation, architectural innovations, and fairness-optimized training objectives. The CLFI metric enables measurement of whether performance disparities reflect inherent task difficulty or model unfairness. Organizations should establish fairness review processes including pre-deployment assessment, participatory data curation, continuous monitoring, and incident response for fairness failures. Regulatory frameworks including the EU AI Act should explicitly address language fairness, as current proposals focus on demographic fairness while omitting language equity protections.

B. Comparison with Existing Approaches

Our approach differs from existing multilingual NLP research in important ways. Prior research on multilingual LLMs focuses on coverage—including as many languages as possible. Our research prioritizes fairness—ensuring equitable performance for included languages. Coverage without fairness is problematic: a model performing poorly on languages it purports to support creates an illusion of inclusion. CLFI extends fairness assessment to language-group performance parity in complex generation and understanding tasks. We systematize participation as integral to fairness assessment rather than supplementary, and explicitly incorporate cultural context analysis recognizing that fairness requires equitable representation and cultural respect.

C. Limitations and Challenges

Our research encounters several limitations: (1) Scope Constraints: Evaluation covers 50 languages and 5 LLM systems; comprehensive assessment of 7,000+ world languages is infeasible. (2) Benchmark Limitations: Existing benchmarks may not measure capabilities important for particular language communities. (3) Baseline Definition: CLFI depends on defining baseline language difficulty through monolingual models, which may reflect training data quality rather than inherent task difficulty. (4) Causality Assessment: Our analysis demonstrates correlation between data availability and performance disparities but cannot definitively establish causation. (5) Generalization: Findings based on contemporary LLMs may not generalize to future architectures.

D. Ethical and Societal Implications

Language fairness in LLMs raises fundamental questions about technology, power, and equity. LLM performance disparities effectively create linguistic discrimination, denying speakers equitable access to beneficial technologies. Low-resource languages are disappearing at alarming rates—one language dies every 2–3 weeks—and LLMs offer potential for language documentation and preservation, but only if systems can effectively process these languages. Language fairness challenges reflect and reinforce existing power imbalances between wealthy nations that control AI development and those who depend on the technologies. When speakers of minority languages lack equitable LLM access, they lose voice in digital conversations increasingly mediated by AI.

## VI. CONCLUSION

This research provides a comprehensive examination of fairness in multilingual LLMs, advancing understanding of this critical challenge and proposing solutions. We systematically map current language fairness in LLMs, documenting 30–45% performance disparities for low-resource languages. We introduce CLFI (Cross-Lingual Fairness Index), extending fairness frameworks from retrieval tasks to LLM generation and understanding. We evaluate the effectiveness of data augmentation, architectural approaches, and participatory design, with regional multilingual models emerging as optimal (CLFI: 78–82). We demonstrate that fairness encompasses not only accuracy parity but equitable representation, knowledge coverage, and cultural respect. We synthesize findings into a practical framework for equitable LLM development emphasizing data curation, participatory involvement, fairness-aware training objectives, and continuous fairness monitoring.

## ACKNOWLEDGMENT

The authors thank the Masakhane community, IndicNLP Consortium, and participating native speaker evaluators for their contributions to this research. This work was conducted at the Department of Computer Science and Engineering, Parul University, Vadodara, Gujarat, India.

## REFERENCES

[1] T. Brown et al., “Language models are few-shot learners,” Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.

[2] OpenAI, “GPT-4 technical report,” arXiv:2303.08774, 2024.

[3] J. N. Pava, A. Singh, and A. Krishnan, “Mind the (Language) Gap: Multilingual LLM performance disparities and equity,” Stanford HAI Report, 2025.

[4] Ethnologue, “Languages of the World,” 27th ed., SIL International, 2024.

[5] UNESCO, “Recommendation on the Ethics of Artificial Intelligence,” Doc. 41 C/Res.88, 2021.

[6] W. Nekoto et al., “Participatory Research for Low-Resourced Machine Translation,” Proc. EMNLP, pp. 61–72, 2020.

[7] P. Joshi et al., “The State and Fate of Linguistic Diversity and Multilingualism in the Age of Deep Learning,” Proc. ACL, pp. 6594–6613, 2021.

[8] A. Conneau et al., “XLM-R: Unsupervised Cross-lingual Representation Learning at Scale,” Proc. ACL, pp. 8440–8451, 2023.

[9] K. Papineni et al., "BLEU: a method for automatic evaluation of machine translation," Proc. ACL, pp. 311–318, 2002.

[10] T. Bolukbasi et al., "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," Proc. NIPS, pp. 4349–4357, 2016.

[11] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," Computational Linguistics, vol. 29, no. 1, pp. 19–51, 2003.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," Proc. NIPS, pp. 3104–3112, 2014.

[13] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. NAACL-HLT, pp. 4171–4186, 2019.

[14] L. Xue et al., "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," Proc. NAACL-HLT, pp. 483–498, 2021.

[15] Masakhane Community, "Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages," Proc. EMNLP, 2020.

[16] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," California Law Review, vol. 104, pp. 671–732, 2016.

[17] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," Proc. FAT*, pp. 77–91, 2018.

[18] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," Proc. NIPS, pp. 3315–3323, 2016.

[19] L. Su et al., "Fairness Measures for Multilingual NLP," arXiv:2106.15596, 2021.

[20] A. Bakarov, "A Survey on Bias and Fairness in Machine Learning," arXiv:2908.04913, 2021.

[21] BigScience Workshop, "BLOOM: A 176B-Parameter Open-Access Multilingual Model," arXiv:2211.05100, 2022.

[22] Google, "Gemini: A Family of Highly Capable Multimodal Models," arXiv:2312.11805, 2024.

[23] Y. Zhang and B. Yang, "A Survey of Multilingual Neural Machine Translation," ACM Computing Surveys, vol. 53, no. 5, pp. 1–38, 2021.

[24] Meta, "No Language Left Behind: Scaling Human-Centered Machine Translation," Proc. EMNLP, pp. 7441–7456, 2024.

[25] N. Goyal et al., "The FLORES Evaluation Datasets for Low-Resource and Multilingual Machine Translation," Proc. EMNLP, pp. 7889–7900, 2021.

[26] E. Yang, J. D. Choi, and J. Callan, "PEER: A Probabilistic Measure of Language Fairness in Information Retrieval," Proc. SIGIR, pp. 1458–1468, 2024.

[27] P. Das et al., "Equality of Opportunity in Machine Translation," Proc. EMNLP, pp. 8398–8410, 2021.

[28] K. Sentiweba et al., "Cultural Fairness in Multilingual Text Generation," arXiv:2309.04521, 2023.

[29] A. Vania et al., "The State of Multilingual NLP: A Survey of Benchmarks, Datasets, and Metrics," ACM Computing Surveys, vol. 56, no. 8, pp. 1–47, 2024.

[30] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," Proc. ACL, pp. 86–96, 2016.