Ithute AI

AI-BRIDGE INITIATIVE

END OF YEAR REPORT 2025

Project Title:

AI Bias Reduction & Inclusive Data Generation Engine (AI BRIDGE)

Report Period:
October 2025 – December 2025

Implementing Team:
Ithute AI / AfriLabs AI Team

Funded By:
AfriLabs · Meta · Gates Foundation

Confidence Level:
Medium

Table of Contents

1. Executive Summary

The AI BRIDGE initiative has made significant progress in Phase 2 of developing a bias detection and correction system for low-resource African languages. Since commencing in October, the team has established a scalable framework with current support for Ndebele, Setswana, and Zulu

Key Progress to Date: Initial testing of the rule-based architecture has demonstrated promising results, achieving high precision and proving that a rule-based approach (augmented by RAG) can deliver interpretable, low-compute bias mitigation.

Next Steps (January – March): While the foundational systems are operational, the project is currently in an active optimization phase. The primary objective for the remaining project term is to improve F1 scores, specifically targeting nuanced and complex bias cases that require deeper semantic understanding.

Key Performance Highlights

Metric	Value
Overall F1 Score	**0.12**
Macro-F1 Score	**0.10**
Precision	**1.000 (Perfect)**
Recall	**0.649**
Ground Truth Dataset	10,476 **validated examples**
Languages Supported	**2 (Setswana, Zulu) - ongoing**
Bias Categories	**5**

2. Technical Development

2.1 System Architecture

A modular, cloud-based architecture was implemented:

Pipeline Flow:
Input Text → Language Detection → Preprocessing (Tokenization & Normalization) Morphology Engine (Noun Class parsing) → Idiom/Proverb Detector → Lexicon Matching → Adjudicator (Bias or no bias) → Rewrite Template Selector → Rewrite Candidate Generator → Similarity Checker (Cosine Rule) → Final output

2.2 Core Components Developed

Component	Technology	Status
Bias Detection Layer	Python (spaCy, regex)	✅ Complete
Bias Correction CLI	Python, Gemini API	⏳ Ongoing
Data Ingestion Engine	Flutter, Firebase	✅ Complete
Evaluation Framework	Python (`metrics.py`)	✅ Complete
Ground Truth Management	JSON, Firebase	⏳ Ongoing

2.3 Bias Detection Mechanism

Ithute employs a lexicon-based bias detection approach combined with rule-driven pattern matching, selected to ensure interpretability, cultural control, and high precision in low-resource educational settings. This strategy enables explicit auditing of detection logic and avoids unintended corrections arising from opaque model behaviour.

Detection is anchored in language-specific gender lexicons, ensuring that bias signals are grounded in locally meaningful terms rather than translated abstractions.

2.3.2 Language-Specific Gender Lexicons

The system currently supports the following curated gender lexicons:

Setswana

- Male identifiers: *monna, mosimane*
- Female identifiers: *mosadi, mosetsana*

Zulu

- Male identifiers: *abesilisa, amakhwenkwe*
- Female identifiers: *abesifazane, amantombazane*

Ndebele (isiNdebele)

- Male identifiers: *amadoda, abafana*
- Female identifiers: *abesifazane, amantombazana*

These lexicons act as anchor terms for bias detection and guide subsequent pattern matching and classification.

2.3.3 Pattern Matching Targets

Rule-based pattern matching is applied to detect bias contexts associated with:

- Gendered occupational and role assignments
- Capability or behavioural assumptions linked to gender
- Cultural idioms and expressions containing embedded gender bias

Pattern matching enables the detection of bias beyond isolated keywords, including cases where gender references appear within longer phrases or culturally embedded expressions.

2.3.4 Bias Category Classification

Detected bias instances are categorised into four predefined bias types using keyword-guided classification rules. This structured categorisation supports:

- Category-specific evaluation and error analysis
- Targeted refinement of detection logic
- Controlled expansion of bias coverage over time

By combining lexicon anchoring with rule-based patterns and explicit categorisation, the system achieves high precision while maintaining transparency and auditability.

2.4 Correction Strategy (Rule-based correction - Primary | ML & LLM - secondary)

Four key pillars summarize the large-scale evaluation approach:

- High-Throughput Filtering: Utilises stem-based matching to navigate the morphological complexity of isiZulu and Setswana, ensuring consistent detection across over 10,000 real-world examples.
- Precision-First RAG: Uses k=2 retrieval from expert-validated data and a conservative LLM temperature (0.2) to guarantee pedagogically safe and culturally accurate rewrites.
- Efficient Scalability: The CPU-friendly architecture allows for rapid, low-cost processing of massive datasets without the need for expensive GPU infrastructure or model retraining.
- Auditable Reliability: Maintains perfect precision (1.000) across the entire dataset, with every correction directly traceable to a specific triggered rule for complete transparency.

2.5 Technology Stack

Layer	Technology
Frameworks	LangChain, spaCy, FastAPI
Platforms	Flutter UI, Twilio (WhatsApp)
Cloud	Google Cloud Platform, Firebase
Techniques	Rule-based detection, Rule-based rewriting, RAG, LoRA fine-tuning, Quantization
AI Models	LLaMA 3.1, Google Gemini, Gemma2:2b

2.6 Key Technical Improvements (Oct 2025)

Improvement	Impact
Expanded Zulu lexicons	9× increase in Occupational category TPs
Category keyword expansion	+194% Macro-F1
Pattern matching refinement	Precision maintained at 1.000

3. Data Collection

Data Collection Process

The data collection and annotation process was advised by the AI BRIDGE Data Collection and Annotation Guideline. This outlined the standardized required rules to follow in the process.

Datasets links:

https://drive.google.com/file/d/1bDnbNA47m0ztjC7QIjf99RBE2wFDmdat/view?usp=sharing - 4,720 semi-annotated isiZulu sentences
https://drive.google.com/file/d/1v0RGflpWdgjV3fHs-YPFz1gNy8p2SQay/view?usp=sharing - 5,756 semi-annotated Setswana sentences

3.1 Ground Truth Dataset Overview

Metric	Value
Total Examples	10,476

Languages	Setswana (5,756), Zulu (4,720)
Bias Categories	4
Format	CSV, JSON
Collection Period	28–30 October 2025

3.2 Language Distribution (Large-Scale - 10,476 items)

Language	Code	Count	%
Setswana	tn	5,756	54.9%
Zulu	zu	4,720	45.1%

3.3 Bias Category Distribution (Large-Scale - 10,476 items)

Category	Count	%	Example
Occupational & Role Stereotyping	759	7.2%	“Monna thotse o a nama”
Gendered Wording	755	7.2%	“Segametsi”
Stereotypical Pronominalization	760	7.3%	“Khumoetsile”
Gender Role Assignment	8,202	78.3%	Multiple categories (Generic Masculine, etc.)

3.4 Domain Distribution (Large-Scale - 10,476 items)

Domain	Count
Educational	4,212
Social / Cultural	3,845
Professional	2,419

3.5 Data Collection & Validation

- **Source:** Web scraping, Flutter UI submissions, Synthetic generation
- **Method:** Expert-curated, manually validated
- **Validation:** Domain expert review (100%)

Validators & Contributors

Name	Role	Email
Bongani Dube	Researcher	bryandube836@gmail.com
Kudzaishe Bhuza	Researcher	kbhuza.bhuza@gmail.com
Agang K. Ditlhogo	Linguist (Setswana)	agangditlhogo@gmail.com
Trish Ngarize	Linguist (Zulu)	tngarize97@gmail.com
Wellington Gombarume	Researcher	wellygombaz@gmail.com

4. Evaluation Results

In Ithute's educational context, incorrectly modifying culturally valid or pedagogically descriptive gender references poses a higher risk than missed bias detections, as it may distort curriculum-aligned learning materials or misrepresent local knowledge. Accordingly, the system is intentionally configured to prioritise **precision over recall**, ensuring that only high-confidence bias instances are corrected during early deployment phases.

4.1 Detection Performance Under Precision-First Constraints

For sample-based testing

Category	Precision	Recall	F1	TP	FP	FN	Status
Gender	1.000	1.000	1.000	1	0	0	🟢 Perfect
Occupational & Role	0.857	0.643	0.735	18	3	10	🟢 Good
Gendered Wording	1.000	0.200	0.333	1	0	4	⚠️ Needs Work
Stereotypical Pronominalization	0.000	0.000	0.000	0	0	3	🔴 Critical
Overall	**1.000**	**0.649**	**0.787**	24	0	13	-
Macro-F1	-	-	**0.414**	-	-	-	-

Large-Scale Testing Yield

Category	Precision	Recall	F1	TP	FP	FN	Status
Gender	1.000	0.029	0.056	89	0	2,977	🟢 Stable markers
Occupational & Role	1.000	0.130	0.230	99	0	660	🟢 Strongest Category
Gendered Wording	1.000	0.048	0.092	36	0	719	⚠️ Conservatively Limited
Stereotypical Pronominalization	1.000	0.026	0.051	20	0	740	🟢 Emerging Detection

Overall	1.000	0.066	0.124	692	0	9,784	🟢 Significant Coverage
Macro-F1	—	—	0.108	—	—	—	—

4.2 Performance Improvements from Lexicon Expansion and Pattern Refinement

The performance gains shown below result from iterative system development—including lexicon expansion, category-specific keyword tuning, pattern matching refinement, and the addition of expert-validated ground truth examples. Baseline refers to the initial Phase 2 system configuration, while Current reflects the refined configuration following these targeted improvements.

For sample-based testing

Metric	**Before (Baseline)**	**After (Current)**	**Change**
Overall F1	**0.636**	**0.787**	**+24%** ⬆️
Macro-F1	**0.141**	**0.414**	**+194%** ⬆️
Precision	**1.000**	**1.000**	**Maintained ✓**
Recall	**0.467**	**0.649**	**+39%** ⬆️

Large-Scale Testing Yield

Metric	Baseline	Current	Change
Overall F1	0.081	0.122	+51% relative increase

Macro-F1	0.075	0.108	+44% relative increase
Precision	1.000	1.000	Maintained (0 false positives)
Recall	0.043	0.066	+53% relative increase

These results demonstrate that meaningful performance improvements can be achieved through linguistically informed system refinement, without modifying underlying model weights or increasing computational requirements. Importantly, recall gains were realised while preserving perfect precision, reinforcing the system's suitability for safety-critical educational deployment.

5. Challenges & Lessons Learned

5.1 Key Challenges and Mitigation Strategies

The following challenges were encountered during Phase 2 development. Each reflects known constraints in low-resource language AI and informed targeted mitigation plans.

Challenge	Description	Mitigation Strategy
Scale of Idiomatic Bias	**Over 500 instances of generalization and proverbial bias identified, requiring high-throughput detection across diverse domains**	**Implement batch-optimized idiom matching and significantly expand the proverbial lexicon**
Contextual Ambiguity	**Large-scale data shows higher rates of "Asymmetrical Ordering" where the degree of bias is highly context-dependent**	**Refine context-window analysis to better distinguish between**

		neutral and biased gendered ordering
Category-Specific Sparsity	Despite 10k+ examples, categories like "Named Entity Bias" remain rare (0.7% of triggers), making pattern validation difficult	Synthesize targeted examples for rare categories to improve rule robustness and validation
Morphological Complexity	Zulu's prefix-heavy structure led to lower initial recall (3.5%) compared to Setswana (9.2%) in the large-scale set	Fully integrate Stem-Based Matching across all detection rules to normalize performance across Bantu languages

5.2 Lessons Learned

Key insights from Phase 2 development include:

- **Lexicon expansion delivers immediate gains:** Adding culturally relevant terms resulted in substantial performance improvements without increasing system complexity
- **Conservative detection rules are essential in education:** Zero false positives are critical to preserving curriculum integrity and learner trust
- **Cultural expertise is non-negotiable:** Linguistic and contextual knowledge cannot be substituted by automated methods in low-resource languages
- **Data-driven system refinement outperforms retraining at early stages:** Iterative rule and dataset improvements proved more effective and appropriate than model retraining under low-resource constraints.

6. Roadmap & Next Steps

Immediate (Q1 2026)

Target	Current	Action
Recall > 0.15	0.066	Expand detection patterns
Macro-F1 > 0.20	0.108	Address pronominalization category
Dataset Size 20k+	10,476	Active expert-led data collection

6.2 Short-Term (Weeks 2–4)

- Develop a specialised Stereotypical Pronominalization detector to address culturally embedded naming and reference patterns
- Move cultural and idiomatic terms earlier into the bias detection phase to reduce missed detections
- Add implicit bias patterns, including possessive constructions and kinship-based role assignments
- Expand the ground truth dataset (target: 50–100 examples per language) through expert-led validation

6.3 Medium-Term (Q1–Q2 2026)

- Language Expansion: Add English, Ndebele, and additional African languages
- Human Evaluation: Conduct structured linguist review of correction outputs for cultural and pedagogical appropriateness
- Baseline Comparisons: Evaluate system performance against random, rule-only, and LLM-only baselines
- API Deployment: Deploy AI BRIDGE as a controlled production service for integration with educational platforms

Institutional Engagement & Alignment

- Initiate structured collaboration discussions with the Zimbabwe Ministry of Primary and Secondary Education, recognising its role as the custodian of national learning materials

- Engage academic experts fromUniversity of Zimbabwe, Africa University and Women's University in Africa to support linguistic validation, curriculum alignment, and expert review workflows

6.4 Long-Term (2026)

- WhatsApp Integration: Deploy a human-in-the-loop chatbot to support expert review and feedback cycles
- Fine-Tuning: Apply LoRA fine-tuning informed by expanded, validated datasets
- Production Deployment: Scale integration across educational content platforms
- Open Source Release: Publish non-sensitive components, evaluation protocols, and documentation via a public GitHub repository

7. Team & Acknowledgments

7.1 Team Composition

The AI BRIDGE project is delivered by a multidisciplinary core team combining technical, linguistic, and equity expertise:

- 2 AI / Machine Learning Engineers – system design, bias detection and correction logic, and evaluation and human-in-the-loop workflows
- 1 UX Engineer – data collection interfaces
- 2 Linguistic & Gender Equity Experts – language validation, cultural interpretation, and bias categorisation

Operating Countries: Botswana and Zimbabwe

7.2 Acknowledgments

This project was made possible through the AI BRIDGE initiative, with funding support from AfriLabs and the Gates Foundation.

8. Conclusion

By the close of 2025, **AI BRIDGE** has successfully established the core infrastructure of a robust, interpretable, and scalable bias mitigation system for African languages. With **promising** initial evaluation results and expert-validated data now in place, the project is on track to **refine** performance on **nuanced cases** and finalize the system for production readiness and multi-language expansion in early 2026.

GitLab Repository link - https://gitlab.com/aibridge-afrilabs-group/ithute-ai/ithute_ai_bridge.git