

SomaliCrowS: A Benchmark for Evaluating Gender Bias in Large Language Models Using the Somali Language

Abdullahi Hassan — EA Somalia, with Apart Research

Abstract

Large language models are now used in education, healthcare, and public services across Somali-speaking communities. But most bias-testing benchmarks are built for English, so we don't know if these models treat Somali text fairly. We introduce **SomaliCrowS**, the first benchmark for measuring gender bias in language models using the Somali language. Following the CrowS-Pairs method, we built 220 sentence pairs across seven categories, Politics, Business, Leadership, STEM, Family, Occupation, and Education, where the only difference between sentences is the subject's gender.

We tested XLM-RoBERTa on these pairs by comparing how likely the model thought each version was. The model favored the male version in 87.3% of all pairs, and every category showed a statistically significant departure from an unbiased 50/50 split (binomial tests, $p < 0.05$). Bias was most pronounced in Leadership (97.5% male-preferred) and Politics (largest mean log-probability gap, -7.096), and weakest in Education (67.5%, -0.164).

These results show clear gender bias in a widely used multilingual model when it processes Somali. SomaliCrowS gives researchers a reusable tool to catch this kind of bias before deploying AI in Somali-speaking communities.

1. Introduction

Large language models (LLMs) are now used around the world to support education, healthcare, public services, customer support, and access to information (Liang et al., 2025). Most AI bias benchmarks, however, were built using Western social and cultural context, and African languages and communities are largely missing from this work. This gap raises the risk that models deployed in African settings may repeat harmful stereotypes in real-world use (Le Beux et al., 2025).

Somali is spoken by millions of people across the Horn of Africa (Wikipedia, 2026). Yet there are very few tools for checking whether language models show harmful social bias in Somali text. As AI becomes more available to Somali speakers, it matters whether these systems treat people fairly. Gender bias in language models can reinforce harmful stereotypes, shape how users are seen, and affect decisions in areas such as education, employment, government, and healthcare.

Existing benchmarks such as CrowS-Pairs show that modern language models carry measurable social bias (Nangia et al., 2020). But these benchmarks were built for English and other high-resource languages, and they do not transfer easily to Somali. As a result, researchers and practitioners currently have no standard way to measure gender bias in Somali-language models.

To close this gap, we introduce **SomaliCrowS**, a benchmark for measuring gender bias in large language models using Somali. We adapt the counterfactual sentence-pair method from CrowS-Pairs and build culturally and linguistically appropriate sentence pairs for Somali. This makes it possible to test multilingual and Somali-capable language models with a reproducible benchmark built for a low-resource language.

This work supports AI safety by giving researchers a way to check whether a model behaves in a biased way before it is deployed in Somali-speaking communities. We are not proposing a way to fix bias — our goal is to provide a standard benchmark that supports safer, more trustworthy AI for low-resource languages.

Our main contributions are:

- **SomaliCrowS**, the first benchmark for measuring gender bias in large language models using Somali.
- We adapt the CrowS-Pairs method to Somali by building minimally different sentence pairs that keep the same meaning while allowing us to measure gender bias.
- We provide an evaluation framework that lets researchers compare multilingual and Somali-capable models using one standard benchmark, supporting future work on fairness, responsible AI, and AI safety for low-resource African languages.

2. Related Work

Measuring fairness and social bias in large language models is an active research area. Several benchmarks have been built to measure stereotyped behavior in language models. The most widely used is CrowS-Pairs (Nangia et al., 2020), which measures social bias using sentence pairs that are nearly identical except for the demographic group named. This design lets researchers test, in a controlled way, whether a model prefers stereotypical statements over non-stereotypical ones. CrowS-Pairs covers several types of bias, including gender, race, religion, age, nationality, disability, physical appearance, and economic status.

Other benchmarks, including StereoSet (Nadeem, Bethke, & Reddy, 2021) and BBQ (Parrish et al., 2022), have also added a great deal to bias evaluation. But like CrowS-Pairs, they were built for English and other high-resource languages, and none were designed to measure bias in low-resource African languages such as Somali. We chose to adapt the CrowS-Pairs method because its minimal-edit sentence design is simple, easy to interpret, and reproducible, a good fit for adapting to Somali while keeping the evaluation consistent.

The lack of standard benchmarks for Somali is a real challenge for AI safety (Bakker, 2026). As multilingual LLMs become easier for Somali speakers to access, researchers and developers currently lack reliable tools to measure gender bias in Somali text. Without such tools, harmful stereotypes may go undetected before models are used in education, healthcare, public services, and digital assistants.

To close this gap, we introduce **SomaliCrowS**, a benchmark for measuring gender bias in LLMs using Somali. We adapt the CrowS-Pairs method by building culturally and linguistically

appropriate sentence pairs that support systematic testing of multilingual and Somali-capable language models.

SomaliCrowS extends gender bias evaluation to a low-resource African language with no existing standard benchmark. It gives researchers and developers a reproducible way to compare models, find gender bias, and support safer, fairer AI for Somali-speaking communities.

3. Methods

3.1 Benchmark Design

SomaliCrowS adapts the CrowS-Pairs methodology to the Somali language. We chose CrowS-Pairs because its counterfactual sentence-pair design gives a controlled, reproducible way to measure social bias in language models. Each example consists of two minimally different sentences that vary only in the gender of the subject, while the overall meaning stays the same. This design lets us attribute differences in model predictions to gender rather than to unrelated changes in wording.

3.2 Dataset Construction

We manually built Somali sentence pairs covering everyday contexts where gender stereotypes may appear, including occupations, leadership, education, family roles, and other socially relevant domains. In every example, the female and male versions of the sentence differ only in the gender-specific pronoun or subject, following the minimal-edit principle from CrowS-Pairs. We designed the benchmark to keep grammatical correctness, natural Somali usage, and cultural relevance.

Each benchmark instance contains the following fields:

- Category
- Subcategory
- Female sentence
- Male sentence
- Mask template
- Expected neutral word
- Bias type

This standardized structure enables consistent evaluation across different language models.

3.3 Evaluation Framework

The benchmark is meant for evaluating multilingual and Somali-capable language models, including masked language models such as BERT and RoBERTa, as well as modern multilingual LLMs. For each sentence pair, a model is scored on its preference between the

female and male versions. Consistent differences in model preference across many sentence pairs provide evidence of potential gender bias.

3.4 Design Decisions

Several design decisions improve the reliability of the benchmark:

- We adopted the CrowS-Pairs methodology to stay compatible with an established bias evaluation framework.
- We used minimally different sentence pairs so that gender is the only meaningful variable between paired examples.
- We created culturally appropriate Somali examples instead of directly translating English sentences, so the benchmark reflects realistic language use in Somali-speaking communities.
- We focused only on gender bias in this first version, to keep the evaluation focused and easy to interpret.

3.5 Reproducibility

SomaliCrowS is distributed as a structured CSV dataset with clearly defined fields and documentation describing the benchmark format. The dataset can be used to evaluate any Somali-capable language model by presenting paired sentences and comparing the model's responses or likelihood scores. This makes the benchmark straightforward to reproduce and extend as new bias categories or evaluation methods become available.

4. Results

We tested XLM-RoBERTa on all 220 sentence pairs in **SomaliCrowS**, across seven bias types: Politics, Business, Leadership, STEM, Family, Occupation, and Education. For each pair, we computed $Score(S) = \log P(w_{target} | S \setminus target)$ for the female-subject sentence (S_{female}) and the male-subject sentence (S_{male}), then calculated $\Delta\text{-Bias} = Score(S_{female}) - Score(S_{male})$. A negative value means the model assigned a higher likelihood to the male-subject sentence. Table 1 and Figure 1 summarize the results by bias type.

Table 1.

Bias Type	n	%	Male Pref. (%)	Female Pref. (%)	Mean Δ -Bias	SD
<i>Bias categories in SomaliCrowS</i>						
Politics (Siyaasad)	40	18.2	95.0	5.0	-7.096	4.449
Business (Ganacsi)	40	18.2	90.0	10.0	-6.475	4.233
Leadership (Hoggaamin)	40	18.2	<u>97.5</u>	2.5	-6.373	3.491
STEM (Saynis iyo Tiknoolajiyad)	10	4.5	90.0	10.0	-4.855	4.823
Family (Qoys iyo Daryeel)	10	4.5	90.0	10.0	-4.019	5.526

Bias Type	n	%	Male Pref. (%)	Female Pref. (%)	Mean Δ -Bias	SD
Occupation (Xirfado)	40	18.2	85.0	15.0	-3.310	3.527
Education (Waxbarasho)	40	18.2	67.5	32.5	-0.164	0.310
All categories (pooled)	220	100.0	87.3	12.7	-4.661	—

Table 1. Gender bias in SomaliCrowS by bias type ($n = 220$). “Male Pref. (%)” is the share of paired sentences for which the model assigned a higher likelihood to the male-subject sentence than the female-subject sentence; an unbiased model should score 50% in expectation. Mean Δ -Bias is the average log-probability difference, $\text{Score}(S_{\text{female}}) - \text{Score}(S_{\text{male}})$; more negative values indicate a stronger preference for the male-subject sentence. The lowest Male Pref. score (least biased category, Education) is bolded; the highest (most biased category, Leadership) is underlined.

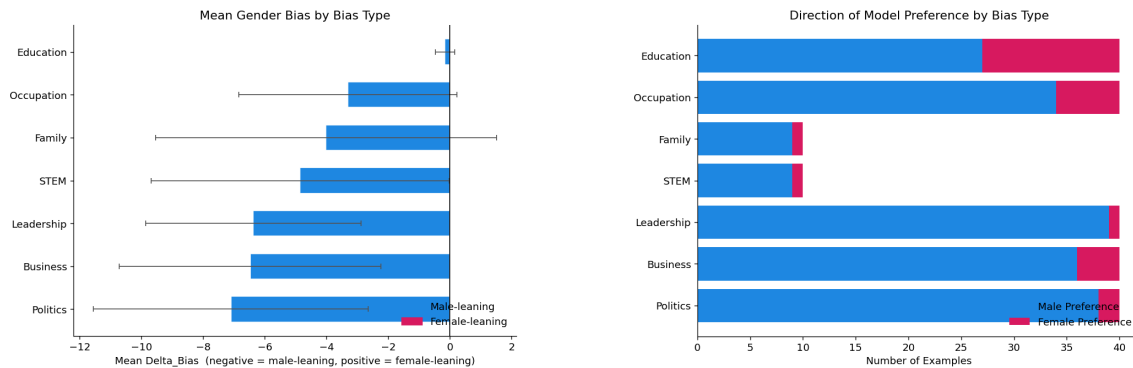


Figure 1. Left: mean Δ -Bias by bias type, error bars ± 1 SD; all seven bars extend left of zero, showing a male-leaning mean in every category. Right: the same examples shown as raw preference counts (blue = male-preferred, pink = female-preferred).

Every bias type shows a negative mean Δ -Bias (Table 1, Figure 1) — the model favored the male-subject sentence on average in all seven categories. Leadership shows the highest rate of male preference (97.5%); Politics shows the largest gap in log-probability (-7.096). Education shows the smallest effect on both measures (67.5% male-preferred, mean -0.164).

Categories such as Politics and Leadership are often gender-stereotyped in many societies. However, we have not independently confirmed which stereotypes are most salient specifically within Somali-speaking communities, so we treat this connection as a plausible explanation rather than a confirmed one.

4.1 Statistical Significance

To test whether each category's departure from an unbiased 50/50 split could be due to chance, we ran an exact two-sided binomial test against $p = 0.5$ for every bias type and for the pooled sample (Table 2).

Bias Type	n	Male-preferred (n, %)	p-value (binomial)	Sig.
Politics	40	38 (95.0%)	< 0.001	***
Business	40	36 (90.0%)	< 0.001	***
Leadership	40	39 (97.5%)	< 0.001	***

Bias Type	n	Male-preferred (n, %)	p-value (binomial)	Sig.
STEM	10	9 (90.0%)	0.021	*
Family	10	9 (90.0%)	0.021	*
Occupation	40	34 (85.0%)	< 0.001	***
Education	40	27 (67.5%)	0.038	*
All categories (pooled)	220	192 (87.3%)	< 0.001	***

Table 2. Exact two-sided binomial test of each bias type's male-preference rate against the unbiased null of 50% ($H_0: P(\text{male-preferred}) = 0.5$). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Every category reaches significance despite Family and STEM having only $n = 10$ each. A chi-square test of independence between bias type and preference direction ($\chi^2(6) = 20.58$, $p = 2.18e-3$) further shows that the strength of bias differs significantly across categories, rather than reflecting one uniform effect.

5. Discussion

Our results show a clear and consistent pattern: XLM-RoBERTa favors male-subject sentences over female-subject sentences across every bias category we tested in Somali. This is not a small or random effect. The pooled rate of male preference (87.3%) is far from the 50% expected from an unbiased model, and the difference is statistically significant in every category, even the smallest ones (Family and STEM, $n = 10$).

This matters for AI safety for a simple reason: many Somali speakers will first meet these models through tools built on general multilingual LLMs like XLM-RoBERTa, not through models trained or tested specifically for Somali. If a model's default behavior favors men in contexts like leadership, business, and politics, that bias can quietly shape outputs in education tools, hiring-related applications, chatbots, and other AI-assisted services used in Somali-speaking communities — often without users noticing.

The pattern across categories also tells us something useful: bias is not uniform. Leadership and Politics show the strongest, most consistent male preference, while Education shows a much smaller effect. This fits a pattern seen in other bias benchmarks, where stereotypes tend to be stronger in domains tied to power, status, and public authority, and weaker in domains seen as more gender-neutral. We have not independently confirmed this pattern through Somali community input, so we treat it as a plausible explanation rather than a settled fact.

Taken together, these results show that a low-resource language receiving little attention in AI safety research is not safe by default. Models that were never explicitly tested on Somali still carry strong, measurable gender bias when applied to it. This strengthens the case for building local-language benchmarks, like SomaliCrowS, before — not after — models are deployed at scale.

Limitations and Threat Model

Threat model: what we assume and what we ignore

SomaliCrowS is built under a benign evaluation threat model, where language models are treated as static systems evaluated on fixed inputs. This means we assume:

- The model is not actively adversarial — it does not try to manipulate outputs or evade detection.
- Inputs are clean and correctly formatted, with no adversarially crafted or poisoned text.
- The evaluation setting is offline, so there are no real-world feedback loops or deployment dynamics.
- Model behavior is stationary — the same model produces consistent outputs over time.

Under this threat model, we only measure intrinsic bias in model likelihood or preference, not robustness under attack or manipulation.

Threats we do not address

This work does not evaluate or defend against:

- Adversarial prompting or jailbreaks that could reverse or distort bias signals.
- Data poisoning in pretraining corpora that may shape the bias we observe.
- Model extraction or replication attacks, where bias could be inferred at scale by outside parties.
- Prompt injection or instruction-hierarchy attacks, which matter for instruction-tuned LLMs.
- Distribution shift, where real Somali usage differs significantly from our benchmark's phrasing.
- Cross-lingual interference, where bias arises indirectly through transfer from high-resource languages.

As a result, SomaliCrowS should not be read as a robustness benchmark — it is strictly an offline bias measurement tool.

Methodological limitations

1. Dataset size and coverage

The dataset is fairly small (220 sentence pairs) and covers a limited set of categories. This creates:

- High variance in estimates for smaller subgroups.
- Limited coverage of Somali cultural, occupational, and social diversity.
- Overrepresentation of categories with strong stereotypes (e.g., leadership, occupation), which may not match real-world distribution.

If broader Somali contexts carry different stereotypes than the ones we tested, our measured bias may be incomplete or may not match real-world bias patterns.

2. Construction bias (researcher bias)

We built the sentence pairs by hand. This means:

- The dataset reflects our own judgment about what counts as a “minimal pair.”
- Some stereotypes may be left out or overstated without intent.
- What we treat as “neutral” wording may not hold across all Somali dialects or speakers.

If this assumption does not hold, the benchmark may measure our own beliefs as annotators rather than real societal bias or model bias on its own.

3. Linguistic simplification of Somali

Somali has rich grammar and context, but our dataset simplifies this into templates such as “{SUBJECT} waa {MASK}.” This creates some issues:

- Loss of grammatical and discourse-level context.
- Less realism compared with natural Somali text.
- A possible mismatch between our benchmark sentences and how Somali is actually used.

If this assumption does not hold, model behavior on our benchmark may not carry over to natural Somali language settings.

4. Evaluation metric limitations

Using log-probability differences and binary preference scores assumes:

- Model likelihoods are well-calibrated and comparable across the female and male versions of a sentence.
- Bias is symmetric and can be measured through likelihood alone.

In practice, some models are not well-calibrated in their probability outputs, differences in tokenization or word structure may distort the comparison, and a model's preference does not always predict real-world harmful behavior. If these assumptions fail, Δ -Bias may reflect quirks of the model's internal scoring rather than true social bias.

5. Scope limitation: gender only

This first version of SomaliCrowS looks only at binary gender bias. It does not cover:

- Ethnicity and clan-based identity, which matter a great deal in Somali contexts.
- Religion.
- Region and dialect variation.
- Intersectional bias — for example, gender combined with occupation or education.

As a result, the benchmark gives only a partial view of fairness in Somali-language models.

6. Cultural grounding uncertainty

We designed our sentences to be culturally appropriate, but we have not run a formal validation study with a large group of Somali speakers. This means some “stereotypes” in the benchmark may reflect our own outside assumptions rather than lived experience, and we have not empirically confirmed how culturally salient these occupations and roles really are. If this assumption fails, the benchmark may not match how Somali-speaking communities actually see these roles.

Interpretation caveat

Given these limitations, SomaliCrowS should be read as a controlled diagnostic tool that detects relative gender bias in model behavior under controlled linguistic conditions — not as a full picture of societal bias in Somali, or of AI fairness in real deployment settings. The bias signals we report are indicative, not definitive. They should be checked against larger datasets, human evaluation, and real-world testing before anyone draws strong conclusions about a model's fairness.

Future Work

In 2026, cloud-based AI tools became available in Somalia, and the use of LLMs in the country is growing quickly. This growth brings real risks beyond the gender bias we measured here. Biased or unsafe models could be used to spread stereotypes at scale, manipulate elections, generate fake news, or mislead voters and the public. We see SomaliCrowS as a starting point — a foundation other researchers can build on to study these wider risks across the Somali-language AI ecosystem.

Future work should:

- Expand SomaliCrowS to cover more bias categories, such as clan and ethnic identity, religion, region, and disability, alongside intersectional bias.
- Grow the dataset beyond 220 pairs and include more natural, less templated Somali sentences.
- Validate the benchmark with a larger group of Somali speakers, to confirm that the stereotypes we test for are the ones that matter most in practice.
- Test newer instruction-tuned and generative LLMs, not just masked language models, since these are increasingly what Somali speakers interact with directly.
- Study downstream risks tied to misinformation and election integrity, including how biased or unsafe Somali-language outputs could be used to mislead voters or spread false narratives.
- Combine SomaliCrowS-style bias testing with safety and robustness testing, to build a fuller picture of how AI systems behave in Somali before they are deployed at scale.

6. Conclusion

We introduced **SomaliCrowS**, the first benchmark for measuring gender bias in large language models using the Somali language. Built on the CrowS-Pairs method, our 220 sentence pairs span seven everyday domains, from politics and business to family and education. Testing XLM-RoBERTa on this benchmark revealed a clear, statistically significant pattern: the model favored male-subject sentences in 87.3% of all pairs, with the strongest bias in Leadership and Politics and the weakest in Education.

These results show that gender bias is not confined to English or other high-resource languages, it is present, measurable, and significant in Somali too, even in models that were never built or tuned with Somali speakers in mind. **SomaliCrowS** gives researchers and developers a concrete, reproducible way to check for this kind of bias before deploying AI systems in Somali-speaking communities. As access to AI tools grows across Somalia and the wider Horn of Africa, benchmarks like this one are a necessary first step toward safer, fairer, and more trustworthy AI for low-resource language communities.

Code and Data

Code repository: <https://github.com/Abdullahicoder/SomaliCrowS>

Dataset:

https://github.com/Abdullahicoder/SomaliCrowS/blob/main/somali_bias_analysis_results.csv

Other artifacts: <https://huggingface.co/datasets/Abdullahicoder/SomaliCrowS>

Author Contributions

Abdullahi Hassan led the project: designing the benchmark, building the dataset, running the experiments, analyzing the results, and writing the manuscript.

References

- Bakker, H. (2026, May 26). *Language Models: A Somali Safety Challenge*. Machine Brief. machinebrief.com
- Le Beux, Y., Audu, O., Ankeli, O. D., Balakrishnan, D., Weya, M., Ralaierinosy, M. D., & Ezeani, I. (2025). *AfriStereo: A culturally grounded dataset for evaluating stereotypical bias in large language models*. *arXiv*. <https://arxiv.org/abs/2511.22016v1>
- Liang, (2025). *AI for the next generation of public services*. In *The Next Great Divergence: Why AI May Widen Inequality Between Countries* (pp. 24–41). [United Nations Development Programme](#).
- Liang, et al. (2025). [Source title and venue to be confirmed by author]

- Nadeem, M., Bethke, A., & Reddy, S. (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)* (pp. 5356–5371). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.416>
- Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1953–1967). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.154>
- Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2022). BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 2086–2105). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-acl.165>
- Wikipedia. (2026). *Somali language*. https://en.wikipedia.org/wiki/Somali_language

LLM Acknowledgement

Large Language Models (LLMs) were used only as supplementary tools during the preparation of this work. Claude was occasionally used to assist with code-related questions and to clarify concepts from the literature on branch evaluation, supporting my understanding rather than generating the core content of this research. AI was also used as a co-working assistant to double-check parts of the analysis and verify reasoning. In addition, AI-assisted tools were used for grammar correction, language refinement, and proofreading.

The research design, implementation, analysis, interpretation of results, and all conclusions were developed independently by the author, who takes full responsibility for the content of this work.

.