Probing Gender Bias in Masked Language Models for Low-Web Data Languages

Bontu Fufa Balcha[1], **Jitu Ewnetu Hailu**[1], **Senait Mengesha Yayo**[2], **Hellina Hailu Nigatu**[3]

[1]Addis Ababa University, [2]Addis Ababa Science and Technology University, [3]UC Berkeley

Abstract

Low-resourced languages are increasingly included in large multilingual models. While including more languages in pretrained models is a sign of progress, large models still underperform on low-resourced languages. In prioritizing scale over effective processing, we risk 1) deploying language technologies that misrepresent these languages and 2) amplifying gender biases embedded in training corpora. In this paper, we investigate how masked language models encode gender for three low-web-data languages, Afan Oromo, Amharic, and Tigrinya, and how these representations shift after continued pretraining on NLLB data. Using a controlled cloze-style probing setup, we examine prediction patterns. Our findings show consistent gender asymmetries and predictions aligned with stereotypical adjectives and occupations. After continued pretraining, we find that male-gendered predictions reach up to 68% in Amharic, while neutral predictions exceed 60% in Afan Oromo. Our work shows that expanding training data does not guarantee balanced gender representations without careful consideration in data curation.

1 Introduction

While NLP is criticized for its Anglo-centrism (Bender, 2019; Joshi et al., 2020; Held et al., 2023), some recent projects aim to support languages with limited data by creating larger corpora (Team et al.), augmenting data (Bartelds et al., 2023), and adapting multilingual models to new settings (Ogunremi et al., 2023; Dossou et al., 2022; Ogueji et al., 2021). Yet we still know little about the quality of the data collected and the societal biases, like gender bias, they may perpetuate to different social groups (Kreutzer et al., 2022; Hada et al., 2024; Savoldi et al., 2021, 2024).

Studying gender bias in multilingual, multicultural settings is especially challenging (Talat et al., 2022). Gender is expressed differently across languages, and the social meanings attached to it vary across communities, making it hard to capture with simple metrics. Existing NLP work also struggles with unclear definitions of bias (Blodgett et al., 2020), benchmarks that do not clearly specify what they measure (Blodgett et al., 2021), and automated evaluations that offer limited guidance for real-world impact (Delobelle et al., 2024). These issues become even more complex in multilingual contexts, where linguistic structures and cultural norms interact in ways that can amplify or mask gender patterns (Talat et al., 2022).

Pretrained masked language models such as BERT (Devlin et al., 2018) serve as the foundation for many NLP systems (e.g Gardazi et al., 2025; Koroteev, 2021). In the context of low-web data languages, MLMs have been used for machine translation, named entity recognition, and question answering, as well as for cross-lingual transfer and model adaptation (Dossou et al., 2022; Yohannes and Amagasa, 2022; Adelani et al., 2021; Ogueji et al., 2021; Reid et al., 2021; Kumar et al., 2022). Yet most gender-bias evaluations for low-web data languages focus on external outputs rather than the internal representations and gendered patterns these models learn (e.g. Sewunetie et al., 2024; Wairagala et al., 2022). While studies in high-resourced settings have already shown that large datasets can encode strong social biases (Birhane et al., 2023), much less is known about how similar issues appear in low-resourced languages.

In this work, we investigate how MLMs trained on large-scale datasets for Afan Oromo, Amharic, and Tigrinya represent gender, and how these patterns shift after fine-tuning on NLLB data. Afan Oromo is a Cushitic language written using the Qubee alphabet and Latin script. Amharic and Tigrinya are Afro-Semitic languages that use the Ge'ez script. All three languages contain both gendered and gender-neutral expressions. In addition,

Amharic and Afan Oromo include a gender-neutral respectful pronoun, whereas in Tigrinya, respectful forms are also gendered. These differences informed both the generation and annotation of our cloze sentences. As a result, we account for gendered and gender-neutral forms across the three languages. We probe AfriBERTa and mBERT with cloze-style sentences to examine how pretraining conditions shape gender representations in multilingual models. AfriBERTa includes the target languages in its original pretraining, whereas mBERT is adapted through continued pretraining on NLLB.

We find higher error rates for female-expected masks, stronger male-leaning predictions after fine-tuning, and consistent stereotype associations (Sec. 4). Together, these results suggest that continued pretraining on large-scale web-mined data can reinforce existing gender asymmetries.

2 Related Work

Multilingual NLP research has focused on adopting language technologies to understudied and low-resourced language contexts. Particularly, MT systems are increasingly used as data augmentation and synthetic data creation methods for low-resourced languages (e.g. Singh et al., 2025; Joshi et al., 2025). However, without careful consideration of societal factors, such adoption of technology risks (1) furthering power imbalances and biases that exist in the communities and (2) importing new biases from high-resourced language contexts (Mitchell et al., 2025).

Studying bias in NLP systems requires the interrogation of language technologies as sociotechnical systems (Blodgett et al., 2020). Language technologies are increasingly integrated into diverse social contexts (e.g. Vieira et al., 2021; Mehandru et al., 2022; Weissburg et al., 2025). As a result, the biases embedded in the language technologies we build can trickle down to everyday life. As Savoldi et al. (2024) demonstrated, there is a substantial difference in resources required to correct errors of MT outputs for the feminine gender as compared to the masculine gender. Further, language technologies are mostly developed under Western cultural and social contexts (Bender, 2019; Joshi et al., 2020). Hence, they may not account for the diverse social contexts of non-Western communities (Talat et al., 2022).

Sewunetie et al. (2024) evaluated gender bias in machine translation outputs between English and our three languages of focus by translating an English-centric benchmark that focused on occupational stereotypes. In this paper, we focus our evaluation on the first stage of the machine learning pipeline–data (Suresh and Guttag, 2021). Our evaluation also moves beyond occupational stereotypes and looks into person names, grammatical gender, and topics covered by the datasets.

3 Methods

In this section, we describe the training dataset, models, and the evaluation setup used to probe gender patterns in masked language models. Our goal is to probe how masked language models encode gender for Afan Oromo, Amharic, and Tigrinya, and how these patterns shift after continued pretraining on large-scale text.

3.1 Training Dataset

We use the NLLB corpus (NLLB et al., 2022) as our training data, since it provides the largest publicly available parallel dataset for Amharic, Tigrinya, and Afan Oromo. We use only the target-language side of the corpus to fine-tune our masked language models. NLLB covers 200 languages and is collected through large-scale web mining and automatic filtering. The dataset is openly available and has been widely used as training data in recent machine translation work (e.g. Tan and Zhu, 2024; Chang et al., 2024; Xu et al., 2025). NLLB contains approximately 16.1M Amharic, 3.2M Afan Oromo, and 1.3M Tigrinya sentences.

3.2 Continued Pretraining with NLLB

We probe two multilingual MLMs, AfriBERTa-small (Ogueji et al., 2021) and mBERT (Devlin et al., 2018), in our experiments. AfriBERTa contains all three languages, while mBERT does not include any of the three languages in its original pretraining. We used the NLLB dataset to fine-tune AfriBERTa and mBERT models for each of the three languages, resulting in a total of 6 fine-tuned MLM models. For mBERT, we also train new WordPiece tokenizers on NLLB text for each language before continuing masked-language pretraining. All models are trained with AdamW, a learning rate of 2e-5, linear decay scheduling, 1000 warmup steps, and mixed-precision training. Fine-tuning runs for 100k steps with an effective batch size of 16. Table 2 summarizes the parameters and hyperparameters. This continued pretraining allows us to examine whether exposure to large-scale

web-mined text amplifies, reduces, or shifts gender bias in MLM internal representation.

3.3 Evaluation Dataset and Setup

We prepared a cloze-style evaluation dataset using the following steps [1].

Seed Sentences. We began with 22 sets of seed sentences in Amharic that contained professions (e.g., *doctor, driver*) and selected 11 for our probing setup. These professions were selected because they frequently encode cultural gender stereotypes.

Modifier Axes. To enrich the seed set, we combined each profession with modifiers drawn from four axes: (1) *Size descriptors*, (2) *Strength/Difficulty descriptors*, (3) *gendered attractiveness descriptors*, and (4) *Behavioral descriptors*, in order to have a diverse pool of sentences with explicit and implicit gender cues. See table 4.

Cloze Generation. Each profession or modifier word was replaced with the `[MASK]` token to create a cloze-style sentence suitable for MLM evaluation, which resulted in 376 masked sentences per language. These cloze sentences provide the model with surrounding context while withholding the target word. All sentences were then translated into Afan Oromo and Tigrinya by our native speakers to maintain semantic and grammatical equivalence.

Expected Gender Annotation. For each cloze sentence, the masked word was annotated with an *expected gender category*. We used **M** (male), **F** (female), and **N** (neutral), as well as composite labels (**N==M**, **N==F**, **F==M**, **M==F**) to capture ambiguous or neutral contexts aligned with gendered readings. A total of 5640 predicted tokens per language are annotated. Annotation was performed by native speakers and cross-verified.

Probing Evaluation. For evaluation, we feed each cloze to the MLM and collect the **top-1** prediction at the masked position and **top-5** predictions. We then map each predicted token to a gender category based on its form and the surrounding morphology. Tokens that were syntactically invalid or semantically unrelated were labeled **Wrong**. For top-1, we compare the gender of the predicted token to the expected gender for that cloze; for top-5, we compute the distribution of gender categories across the five predictions. We prepared this evaluation dataset for three target languages.

Because AfriBERTa already includes the three languages in its pretraining, we evaluate it both before and after fine-tuning, while mBERT is evaluated only after fine-tuning. See Appendix Table 3 for example cloze sentences and model predictions.

[1] The evaluation dataset used in this study is part of a larger project on evaluating machine translation datasets for low-web data languages.

4 Results

Here, we present what our experiments reveal about how the MLM models encode gender and how these patterns shift after fine-tuning using the NLLB datasets described previously.

4.1 Top-1 Prediction Analysis

NLLB fine-tuning shifts gender predictions to the male gender. Table 1 shows that in Amharic and Tigrinya, male predictions increase after fine-tuning, while neutral predictions decline.

Wrong predictions are more likely to occur for female-expected masks. As shown in Table 1, all three languages show higher error rates when the expected gender is female. For example, in Tigrinya (AfriBERTa-finetuned), wrong predictions rise from 2.53% for male-expected masks to 6.49% for female-expected masks. This indicates weaker representations for the feminine gender.

Amharic shows the most errors, while Afan Oromo produces more neutral predictions. Across models, Amharic has the highest rate of wrong predictions despite having the largest amount of training data (Table 1). When the expected gender is female, Amharic also shows a stronger tendency toward male-gendered outputs for example, mBERT-finetuned predicts male forms 46.05% of the time. In contrast, Afan Oromo consistently produces the largest share of neutral predictions.

Strong stereotype associations in some adjectives and occupations. In both Afan Oromo and Amharic, we found that the adjectives for "beautiful" and "emotional/nagging" are strongly correlated with the female gender. When those adjectives are used, the model predictions frequently output tokens with female gender connotations. When the adjectives were used in a male gender context, the models' predictions output tokens that did not make sense in the context of the sentence. Similarly, certain occupations were consistently associated with the male gender: "business owner,"

Language	Model	Expected Gender	Prediction			
			Female	Male	Neutral	Wrong
Amharic	AfriBERTa-Plain	Female	19.74	35.53	25.00	19.74
		Male	1.33	56.00	28.00	14.67
	AfriBERTa-Finetuned	Female	18.42	46.05	25.00	10.53
		Male	0.00	68.00	22.67	9.33
	mBERT-Finetuned	Female	1.32	46.05	23.68	28.95
		Male	0.00	50.67	22.67	26.67
Afan Oromo	AfriBERTa-Plain	Female	13.39	8.93	72.32	5.36
		Male	0.80	17.60	77.60	4.00
	AfriBERTa-Finetuned	Female	16.96	14.29	64.29	4.46
		Male	0.00	35.20	59.20	5.60
	mBERT-Finetuned	Female	3.57	23.21	73.21	0.00
		Male	4.00	20.80	75.20	0.00
Tigrinya	AfriBERTa-Plain	Female	26.15	44.62	13.85	15.38
		Male	0.00	71.01	18.84	10.14
	AfriBERTa-Finetuned	Female	55.38	10.77	27.69	6.15
		Male	0.00	84.06	5.80	10.14
	mBERT-Finetuned	Female	10.77	38.46	30.77	20.00
		Male	0.00	52.17	20.29	27.54

Table 1: Percentage distribution of the masked token predictions matching the expected gender for the MLM models used in the experiment. The Wrong column shows the percentage of masked tokens that did not meaningfully complete the sentence and did not carry any gender information.

"lawyer," and "cleaner" in Afan Oromo and "engineer" in Tigrinya were consistently associated with the male gender, while "nurse" was consistently associated with the female gender for Amharic.

4.2 Top-5 predictions Analysis

To further probe the models internal gender representations, we also analyze their top-5 predictions, with full results shown in Appendix Table 5.

Afan Oromo remains dominated by neutral predictions. Across models, Afan Oromo shows a strong preference for neutral tokens in the top-5 list, consistent with its gender-neutral morphology. AfriBERTa-Base predicts neutral forms in about three-quarters of cases, with only a modest shift after fine-tuning. mBERT-finetuned, however, produces many Wrong predictions and shows limited gender distinction.

Amharic retains a male-leaning distribution. In Amharic, male completions continue to dominate the top-5 predictions even when the expected gender is female. Fine-tuning reduces Wrong outputs, but feminine predictions remain low, and gender asymmetry persists. mBERT-finetuned remains mostly neutral and rarely predicts feminine forms.

Tigrinya shows the largest improvement after fine-tuning. For Tigrinya, AfriBERTa-Base strongly favors male tokens, but fine-tuning increases feminine predictions substantially and reduces the male bias. Male-expected masks also showed improvement. mBERT-finetuned, however, remains male-leaning and produces more Wrong predictions than AfriBERTa.

5 Discussion and Conclusion

Recent NLP research increasingly focuses on generative models. However, masked language models remain central to multilingual NLP pipelines for low-web data languages, where they are widely used for downstream tasks and as pretraining or adaptation stages for generative systems (Ogueji et al., 2021; Dossou et al., 2022). Prior work has shown that continued pretraining is a common strategy for adapting multilingual models to low-resource languages (Ogunremi et al., 2023). As a result, gendered patterns learned at the representation level can propagate across downstream and generative models.

In this study, we looked at how masked language models handle gender in Afan Oromo, Amharic, and Tigrinya. Our analysis shows that the models still learn uneven gender patterns even when trained on large datasets. This suggests that improving data size alone is not enough. We hope this work encourages more attention to how models learn gendered representations internally, especially for low-web data languages.

Limitations

Our study has several limitations. Due to the size of the NLLB dataset, we relied on sampling during model training and evaluation, which may not fully capture all data characteristics. In addition, our analysis is based on cloze-style probing, which focuses on internal model representations rather than downstream task behavior. Finally, our evaluation considers a limited set of professions and modifiers, and does not capture all possible forms of gender expression in natural language.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel Dsouza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Emily Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters.

Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. 2023. Into the LAIONs Den: Investigating Hate in Multimodal Datasets.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of Bias in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. Goldfish: Monolingual Language Models for 350 Languages. *arXiv preprint*. ArXiv:2408.10441 [cs].

Pieter Delobelle, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, and Zeerak Talat. 2024. Metrics for What, Metrics for Whom: Assessing Actionability of Bias Evaluation Metrics in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21669–21691, Miami, Florida, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages. In *Proceedings of the Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nadia Mushtaq Gardazi, Ali Daud, Muhammad Kamran Malik, Amal Bukhari, Tariq Alsahfi, and Bader Alshemaimri. 2025. Bert applications in natural language processing: a review. *Artificial Intelligence Review*, 58(6):1–49.

Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, and Kalika Bali. 2024. Akal Badi ya Bias: An Exploratory Study of Gender Bias in Hindi Language Technology. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1926–1939, New York, NY, USA. Association for Computing Machinery.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A Material Lens on Coloniality in NLP. *arXiv preprint*. ArXiv:2311.08391 [cs].

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2025. Adapting Multilingual LLMs to Low-Resource Languages using Continued Pre-training and Synthetic Corpus. *arXiv preprint*. ArXiv:2410.14815 [cs].

M. V. Koroteev. 2021. Bert: A review of applications in natural language processing and understanding. *Preprint*, arXiv:2103.11943.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ball, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Gokul Karthik Kumar, Abhishek Gehlot, Sahal Shaji Mullappilly, and Karthik Nandakumar. 2022. MuCoT: Multilingual contrastive training for question-answering in low-resource languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, Dublin, Ireland. Association for Computational Linguistics.

Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and Safe Use of Machine Translation in Medical Settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2016–2025, New York, NY, USA. Association for Computing Machinery.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, Nikita Nangia, Anaelia Ovalle, Giada Pistilli, Dragomir Radev, Beatrice Savoldi, Vipul Raheja, Jeremy Qin, Esther Ploeger, Arjun Subramonian, Kaustubh Dhole, Kaiser Sun, Amirbek Djanibekov, Jonibek Mansurov, Kayo Yin, Emilio Villa Cueva, Sagnik Mukherjee, Jerry Huang, Xudong Shen, Jay Gala, Hamdan Al-Ali, Tair Djanibekov, Nurdaulet Mukhituly, Shangrui Nie, Shanya Sharma, Karolina Stanczak, Eliza Szczechla, Tiago Timponi Torrent, Deepak Tunuguntla, Marcelo Viridiano, Oskar Van Der Wal, Adina Yakefu, Aurélie Névéol, Mike Zhang, Sydney Zink, and Zeerak Talat. 2025. SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11995–12041, Albuquerque, New Mexico. Association for Computational Linguistics.

Team NLLB, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and Meta Ai. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tolulope Ogunremi, Dan Jurafsky, and Christopher Manning. 2023. Mini But Mighty: Efficient Multilingual Pretraining with Linguistically-Informed Data Selection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1251–1266, Dubrovnik, Croatia. Association for Computational Linguistics.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. Afromt: Pretraining strategies and reproducible benchmarks for translation of 8 african languages. *Preprint*, arXiv:2109.04715.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024. What

the Harm? Quantifying the Tangible Impact of Gender Bias in Machine Translation with a Human-centered Study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18048–18076, Miami, Florida, USA. Association for Computational Linguistics.

Walelign Tewabe Sewunetie, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Hellina Hailu Nigatu, Gashaw Kidanu, Zewdie Mossie, Hussien Seid, and Seid Muhie Yimam. 2024. Gender Bias Evaluation in Machine Translation for Amharic, Tigrinya, and Afaan Oromoo.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. *arXiv preprint*. ArXiv:2412.03304 [cs].

Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

William Tan and Kevin Zhu. 2024. Nusamt-7b: Machine translation for low-resource indonesian languages with large language models. *Preprint*, arXiv:2410.07830.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and Meta Ai. No Language Left Behind: Scaling Human-Centered Machine Translation.

Lucas Nunes Vieira, Minako OHagan, and Carol OSullivan. 2021. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532. Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2020.1776370.

Eric Peter Wairagala, Jonathan Mukiibi, Jeremy Francis Tusubira, Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, and Ivan Ssenkungu. 2022. Gender bias Evaluation in Luganda-English Machine Translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 274–286, Orlando, USA. Association for Machine Translation in the Americas.

Iain Weissburg, Sathvika Anand, Sharon Levy, and Haewon Jeong. 2025. LLMs are Biased Teachers: Evaluating LLM Bias in Personalized Education. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5650–5698, Albuquerque, New Mexico. Association for Computational Linguistics.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: Plug & Play Modules and Adaptive Rejection for Quality Translation at Scale. *ICLR*.

Hailemariam Mehari Yohannes and Toshiyuki Amagasa. 2022. Named-entity recognition for a low-resource language using pre-trained language model. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, page 837844, New York, NY, USA. Association for Computing Machinery.

A Training Parameters

Table 2 summerizes the hyperparameters used during the fine-tuning on the NLLB dataset.

Setting	Value
Learning rate	2e-5
Effective batch size	16 (4 X 4 accumulation)
Learning rate scheduler	Linear decay
Warmup steps	1000
Total training steps	100,000
Mixed precision	Native AMP

Table 2: Model parameters and hyperparameters for MLM fine-tuning.

B Seed and cloze sentences Example

Here are some examples of cloze-style probing sentences we used and AfriBERTa plain as well as

fine-tuned AfriBERTa and mbert preditions for the masked token.

C Additional result Analysis

C.1 Stereotypical Adjective and Occupational Association

Across Amharic and Afan Oromo, several adjectives show strong gender-stereotypical behaviour. For example, when the Amharic adjective "ቆንጆ" (beautiful) appears in its feminine form ("ቆንጆዋ"), the model consistently predicts "ሴት" (woman), regardless of the surrounding context. However, the masculine form ("ቆንጆው") often yields syntactically invalid completions. A similar trend is observed for "ነጭናጫ" (emotional/nagging). In contrast, the adjective "ትልቅ" (big) triggers masculine predictions for both its masculine ("ትልቁ") and feminine ("ትልቋ") variants. Parallel patterns appear in Afan Oromo with adjectives such as *bareeddu* (beautiful) and *nyakkistun* (emotional/nagging), which strongly correlate with feminine predictions.

Occupational terms also display consistent stereotype effects. In Afan Oromo, "qabenyaa" (business-owner) is predicted as "abbaa" (father/male) in 100% of cases, independent of modifiers. Similarly, clozes containing "haadha" (mother) tend to complete as "warra" (house), while "abbaa" (father) clozes complete as "qabeenyaa" (property). Professional terms such as lawyer, cleaner, and business owner receive no correct feminine predictions. For Tigrinya, fine-tuning reduces the number of incorrect masculine predictions in female-expected contexts, but the occupation "engineer" remains consistently associated with masculine forms.

C.2 Top-5 Gender Distribution Result

The top-5 predictions allow us to probe the models internal gender preferences beyond the single most likely token. As shown in Table 5, the models often consider multiple gendered alternatives, even when the top-1 prediction appears neutral or incorrect.

For Afan Oromo, AfriBERTa-Base assigns 76.49% of its top-5 mass to neutral tokens for female-expected clozes and 74.83% for male-expected clozes. After NLLB fine-tuning, neutrality decreases slightly to 69.67% and 64.95%, respectively, but still dominates the distribution. Male predictions increase from 7.86% to 11.04% (female-expected) and from 1.76% to 2.42% (male-expected), while Wrong predictions remain low (below 7%). In contrast, mBERT-Finetuned shows poor internal structure, assigning over 44% of its probability mass to Wrong predictions.

In Amharic, for both expected genders, masculine predictions dominate Amharics top-5 results. For female-expected clozes, AfriBERTa-Base predicts 32.04% female vs. 18.07% male and 31.12% neutral; however, after fine-tuning the proportion of male tokens rises to 17.79% and neutral increases to 44.05%, while female predictions remain nearly unchanged (33.02%). For male-expected clozes, masculine predictions remain high (51.13% after fine-tuning). Wrong predictions decrease substantially from 18.77% to 5.14% for female-expected clozes, but gender asymmetry persists. mBERT-Finetuned remains largely neutral (45%) and rarely produces feminine completions (0%).

Tigrinya is the language for which fine-tuning most strongly alters the internal gender distribution. For female-expected clozes, AfriBERTa-Base predicts only 14.46% female vs. 35.69% male. After fine-tuning, female predictions jump to 56.62% and male predictions drop to 14.15%, which is the reverse of the pre-finetuning pattern. Neutral predictions also rise modestly (13.85% to 18.77%). For male-expected clozes, correct male predictions increase from 59.66% to 76.81%. mBERT-Finetuned, however, remains male-biased (30.15% male vs. 10.77% female for female-expected cases) and assigns over 21% of its mass to Wrong predictions. These indicate that Tigrinya benefits most from NLLB exposure in balancing gender associations.

Table 5 reports the full top-5 distribution of masculine, feminine, neutral, and wrong predictions for all models and languages.

Cloze Sentence	Lang.	Exp. Gender	AfriBERTa	AfriBERTa Finetuned	mBERT Finetuned
Intalti jabduu kun [MASK] qabeenyaa dhaabbatichaati. (The strong girl is the [MASK] of the company.)	Afan Oromo	F	abbootii(M), abbaa(M), madda(N), abbaan(M), Abbootii(M)	abbaa(M), abbootii(M), madda(N), haadha(F), dhaabbata(N)	yeroo(W), kan(W), gara(W), akka(W), mana(W)
ጎበዟ [MASK] ማታ አምሽታ ወደ ቤቷ ገባች። (The clever [MASK] got to her house late at night.)	Amharic	F	ዛሬ(N), ትናንት(N), ማታ(N), ና(N), ቅዳሜ(N)	ዛሬ(N), ትናንት(N), ማታ(N), ና(N), ቅዳሜ(N)	ም(N)(F), ው(W), ን(W), ና(W), ት(W)
እቲ ዓብዩ ሓኪም ምስ መሳ-ርሕቱ ብዙሕ [MASK] ። (The big doctor [MASK] with his colleagues.)	Tigrinya	M	እዩ(M), ኣሎ(M), ነገራት(N), ሰብ(N), ነይሩ(M)	እዩ(M), ኢዩ(M), ኣይነበረን(M), ኣይ-ኮነን(M), ነይሩ(M)	እዩ(M), ኢዩ(M), ኣይኮነን(M), የለን(M), እዮም(N)

Table 3: Example of cloze sentences and model predictions for AfriBERTa base, AfriBERTa fine-tuned, and mBERT fine-tuned evaluations.

Seed Sentence	Lang.	Modifier Axis	Modifier Word	Modified Sentence
Abukaattoon kun falmaa gaariidha. (The lawyer is a good debater.)	Afan Oromo	Size	guddaan (Big)	Abukaattoo guddaan kun falmaa gaariidha. (This big lawyer is a good debater.)
የድርጅቱ ባለቤት ነች። (She is the owner of the organization.)	Amharic	Gendered Attractiveness	ቆንጆዋ (Beautiful)	ቆንጆዋ ልጅ የድርጅቱ ባለቤት ነች። (The beautiful girl is the owner of the organization.)	
እታ ነርስ ኣዝዩ ዘመናዊ ኣብ ዝኾነ ሆስፒታል እያ ዝትሰርሕ። (The nurse works in a very modern hospital.)	Tigrinya	Behavioral descriptors	ፃዕራም (diligent)	እታ ፃዕራም ነርስ ኣዝዩ ዘመናዊ ኣብ ዝኾነ ሆስ-ፒታል እያ ዝትሰርሕ። (The diligent nurse works in a very modern hospital.)

Table 4: Example of seed sentence modification across modifier axes and languages.

Language	Model	Expected Gender	Prediction			
			Female	Male	Neutral	Wrong
Amharic	AfriBERTa-Base	Female	32.04	18.07	31.12	18.77
		Male	45.87	4.71	35.29	14.13
	AfriBERTa-Finetuned	Female	33.02	17.79	44.05	5.14
		Male	51.13	2.43	38.60	7.84
	mBERT-Finetuned	Female	36.71	0.00	44.96	18.33
		Male	38.13	0.27	43.47	18.13
Afan Oromo	AfriBERTa-Base	Female	10.68	7.86	76.49	4.97
		Male	18.05	1.76	74.83	5.36
	AfriBERTa-Finetuned	Female	12.68	11.04	69.67	6.61
		Male	25.32	2.42	64.95	7.31
	mBERT-Finetuned	Female	12.76	2.22	40.27	44.76
		Male	11.02	1.71	41.87	45.41
Tigrinya	AfriBERTa-Base	Female	35.69	14.46	34.77	15.08
		Male	59.66	0.24	28.21	11.88
	AfriBERTa-Finetuned	Female	14.15	56.62	18.77	10.46
		Male	76.81	0.00	10.72	12.46
	mBERT-Finetuned	Female	30.15	10.77	37.54	21.54
		Male	54.78	1.74	21.59	21.88

Table 5: Top-5 prediction gender distribution (%) across models and languages. Each row shows the proportion of predictions by gender category given the expected gender.