Beyond Generalization: Evaluating Multilingual LLMs for Yorùbá Animal Health Translation

Godwin Adegbehingbe, Anthony Soronnadi, Ife Adebara ,Olubayo Adekanmbi
Research & Innovation Department
Data Science Nigeria
Lagos, Nigeria
{Godwin,Anthony,Ife,Olubayo}@datasciencenigeria.ai

Abstract

Machine translation (MT) has advanced significantly for high-resource languages, yet specialized domain translation remains a challenge for low-resource languages. This study evaluates the ability of state-of-the-art multilingual models to translate animal health reports from English to Yorùbá, a crucial task for veterinary communication in underserved regions.We curated a dataset of 1,468 parallel sentences and compared multiple MT models in zero shot and fine-tuned settings. Our findings indicate substantial limitations in their ability to generalize to domain-specific translation, with common errors arising from vocabulary mismatch, training data scarcity, and morphological complexity. Fine-tuning improves performance, particularly for the NLLB 3.3B model, but challenges remain in preserving technical accuracy. These results underscore the need for more targeted approaches to multilingual and culturally aware LLMs for African languages.

1 Introduction

Machine translation (MT) has the potential to improve communication in African languages, but most state-of-the-art models underperform in specialized domains. Yorùbá-speaking communities rely on accurate veterinary translations for disease surveillance and livestock health. However, generic MT models struggle with technical terms and tonal complexities. This study evaluates MT models for domain-specific translation, highlighting challenges and improvements through fine-tuning.

2 Related Work

Recent advances in machine translation (MT) have significantly improved low-resource language translation through transfer learning and unsupervised MT techniques. For African languages, particularly Yorùbá, pre-trained multilingual models like mT5 and mBART (Lee et al., 2022)have shown promising results when fine-tuned on Yorùbá data (Adelani et al., 2022). However, challenges persist in domain-specific applications, especially in specialized fields such as animal health, where standardized terminologies are often absent or underdeveloped (Abenet). Existing MT systems such as NLLB and Google Translate frequently produce erroneous translations of technical terms, highlighting the need for domain-specific fine-tuning (Adebara and Abdul-Mageed, 2022).To address data scarcity in low-resource MT systems, researchers have explored various augmentation techniques. Back-translation has shown promise by creating synthetic parallel data from monolingual target-language content(Jauregi Unanue and Piccardi, 2020), though its effectiveness in preserving technical accuracy remains uncertain for domain-specific translations(Baruah and Singh, 2022).Synthetic data generation techniques have been investigated for neural MT (Tonja et al., 2023), while human-in-the-loop strategies incorporating domain experts (Nunes Vieira, 2019) have emerged as crucial approaches for improving translation quality, particularly in specialized domains (Yang et al., 2023). Evaluation of MT systems in specialized domains requires comprehensive assessment approaches that go beyond traditional metrics. While metrics such as BLEU, AfriComet and chrF provide insights into different aspects of translation quality, (Zappatore and Ruggieri, 2023) argue that specialized domains like biomedical MT require tailored evaluation strategies emphasizing terminology accuracy and practical usability. For Yorùbá animal health translation, these metrics collectively offer a multi-faceted assessment framework: BLEU measures n-gram overlap, AfriComet accounts for semantic accuracy in African languages, and chrF captures character-level precision, particularly valuable for morphologically rich languages like Yorùbá.

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 192–194
July 31, 2025 ©2025 Association for Computational Linguistics

3 Dataset and Methodology

We introduce VetYorùbá, a curated corpus of 1,468 English-Yorùbá parallel sentences, sourced from veterinary health reports. Data preprocessing included normalization to handle Yorùbá's tonal orthography. We evaluated multiple MT models, including NLLB 3.3B (Team et al., 2022), AfriTeVa (Jude Ogundepo et al., 2022), and mT0, under zero-shot and fine-tuned conditions. Metrics such as BLEU, chrF, and AfriComet were used to assess translation quality.We collected our data from three primary sources: the World Organisation for Animal Health (WOAH) reports focusing on seven epidemiologically significant diseases in the region: Rabies, Avian Influenza, Newcastle Disease, Foot-and-Mouth Disease (FMD), African Swine Fever (ASF), Bovine Tuberculosis, and Peste des Petits Ruminants (PPR). Food and Agriculture Organization (FAO) documentation covering animal health practices, preventive measures, and outbreak management protocols, selected to enhance the corpus's terminological breadth. Real-time epidemiological data extracted using PADI-Web (Valentin et al., 2020), an event-based surveillance tool that aggregates information from both structured (official reports) and unstructured sources (news articles, social media)(Oladipo et al., 2023). We focused on maintaining a balanced representation across different disease contexts and livestock categories.Veterinarians facilitated data curation, while native speakers of Yorùbá translated the sentences. The translations were then validated by veterinarians fluent in Yorùbá.

Split	Size	TTR (English)	TTR (Yoruba)
Train	1172	0.2243	0.1672
Dev	147	0.4706	0.3629
Test	147	0.4592	0.3485

Table 1: Dataset split and Type-Token Ratio(TTR) for English and Yoruba sentences

4 Results and Discussion

Zero-shot translation yielded poor results in all models, with NLLB 3.3B achieving a BLEU score of 2.9. Fine-tuning improved performance significantly, raising BLEU to 45.89 for NLLB 3.3B and enhancing chrF and AfriComet scores. However, translation errors persisted, particularly in complex veterinary terms and tonal variations. These findings highlight the limitations of general-purpose LLMs in handling domain-specific, low-resource languages.

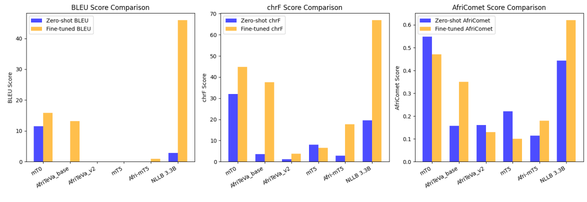

Figure 1: MT Model performance on Yoruba Animal Health Translation

The performance of the machine translation models evaluated was quantified using BLEU (Papineni et al., 2002), chrF (Popović, 2015), and AfriComet (Wang et al., 2024) metrics under both zero-shot and fine-tuned conditions. Overall, fine-tuning on our domain-specific dataset of 1,468 English–Yorùbá sentence pairs resulted in marked improvements across all metrics. In the zero-shot setting, the models generally exhibited low performance, with many struggling to produce coherent translations in the specialized domain of animal health. mT0 achieved a BLEU score of 11.57, while other models such as Afri-mT5 and AfriTeVa_v2 recorded near-zero BLEU scores (0.0003 and 0.005, respectively).Fine-tuning of the models on the curated veterinary dataset significantly improved translation quality. The BLEU score of the mT0 model improved to 15.9, while NLLB 3.3B exhibited the most dramatic gain, rising from 2.9 to 45.89. This improvement was consistently reflected in the chrF scores, with NLLB 3.3B increasing from 19.47 to 66.85. The AfriComet metric further supported these improvements, particularly for the NLLB 3.3B and the AfriTeVa base, whose fine-tuned scores of 62 and 35, respectively, signified better semantic alignment and contextual accuracy in translations.The substantial improvements observed in key models, particularly NLLB 3.3B, confirm that fine-tuning can mitigate the limitations of zero-shot translation (Alabi et al., 2022) and lead to more accurate and reliable translations of technical content in Yorùbá.

5 Conclusion and Future Work

This study underscores the challenges of applying multilingual LLMs to specialized translation tasks in African languages. Although fine-tuning improves performance, key limitations remain, emphasizing the need for tailored approaches integrating linguistic features such as tone and morphology.

Future research would focus on expanding domain-specific corpora and developing African-centric models for technical translation tasks in animal health.

References

T. A. Abenet. Bridging the gap: Legal and medical translation in African indigenous languages. In *Proceedings of ...*

I. Adebara and M. Abdul-Mageed. 2022. Towards afrocentric nlp for african languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

D. I. Adelani, J. O. Alabi, A. Fan, J. Kreutzer, X. Shen, M. Reid, D. Ruiter, D. Klakow, P. Nabende, E. Chang, et al. 2022. A few thousand translations go a long way! Leveraging pre-trained models for African news translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Rupjyoti Baruah and Anil Kumar Singh. 2022. *A Clinical Practice by Machine Translation on Low Resource Languages*. CRC Press eBooks.

I. Jauregi Unanue and M. Piccardi. 2020. Pretrained language models and backtranslation for English-Basque biomedical neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 826–832, Online. Association for Computational Linguistics.

Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. 2022. AfriTeVA: Extending ?small data? pretraining approaches to sequence-to-sequence models. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 126–135, Hybrid. Association for Computational Linguistics.

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D. McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Lucas Nunes Vieira. 2019. *Post-Editing of Machine Translation*, pages 319–335.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. 2023. Better quality pre-training data and t5 models for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore. Association for Computational Linguistics.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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, and Angela Fan. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

A. L. Tonja, O. Kolesnikova, A. Gelbukh, and G. Sidorov. 2023. Low-resource neural machine translation improvement using source-side monolingual data. *Applied Sciences*, 13(1201).

Sarah Valentin, Elena Arsevska, Sylvain Falala, Jocelyn de Goër, Renaud Lancelot, Alizé Mercier, Julien Rabatel, and Mathieu Roche. 2020. Padi-web: A multilingual event-based surveillance system for monitoring animal infectious diseases. *Computers and Electronics in Agriculture*, 169:105163.

Jiayi Wang, David Ifeoluwa Adelani, and Agrawal. 2024. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages. *Preprint*, arXiv:2311.09828.

Xinyi Yang, Runzhe Zhan, Derek F. Wong, Junchao Wu, and Lidia S. Chao. 2023. Human-in-the-loop machine translation with large language model. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 88–98, Macau SAR, China. Asia-Pacific Association for Machine Translation.

M. Zappatore and G. Ruggieri. 2023. Adopting machine translation in the healthcare sector: A methodological multi-criteria review. *Computer Speech & Language*, page 101582.