Rewriting the Narrative: Evaluating Machine Translation in African Languages with Corrected FLORES Data

Itumeleng Moshokoa and Bukhosi Eugene Mpande and Ennis Maphasha
University of Pretoria
Group 24

Abstract

This project examines the impact of dataset corrections on the performance of machine translation (MT) systems for three low-resource African languages: Hausa, Northern Sotho (Sepedi), and isiZulu. While the original FLORES-101 benchmark marked a significant advancement in evaluating MT for under-represented languages, it contained linguistic inaccuracies that could compromise the integrity of model assessments. In response, the Data Science for Social Impact (DSFSI) research group at the University of Pretoria developed corrected versions, collectively known as FLORES Fix for Africa, to improve translation consistency and quality. This study investigates how these corrections affect the evaluation of state-of-the-art MT models, with a focus on translation accuracy and reliability.

1 Introduction

Machine Translation (MT) refers to the automated process of translating text from one natural language to another using computational methods (Goyal et al., 2022a). Over time, MT has evolved significantly, transitioning from rule-based systems to data-driven paradigms(Wang et al., 2022). In particular, neural machine translation (NMT) has emerged as the dominant approach, achieving state-of-the-art results across numerous language pairs (Wang et al., 2022). Despite these advancements, the quality of translation for low-resource languages remains a major challenge due to the limited availability of parallel data, which hampers both model training and evaluation (Costa-Jussà et al., 2022). To address these limitations, benchmark datasets such as FLORES-101 and FLORES-200 (Facebook Low Resource Language Evaluation Sets) have been introduced(Costa-Jussà et al., 2022). These datasets are specifically designed to assess the performance of MT systems, with a focus on improving evaluation for under-represented languages(Costa-Jussà et al., 2022).

As noted in (Abdulmumin et al., 2024b), the original FLORES dataset represented a significant advancement by incorporating under-represented languages, thereby contributing to more inclusive machine translation research. However, despite this progress, the dataset exhibited various inconsistencies and inaccuracies that could compromise the validity of evaluations in downstream Natural Language Processing (NLP) tasks(Abdulmumin et al., 2024b). To mitigate these issues, targeted corrections were implemented for three African languages(Abdulmumin et al., 2024b). These revisions aimed to enhance the linguistic accuracy and consistency of the dataset, thereby enabling more reliable and precise evaluations of machine translation quality(Abdulmumin et al., 2024b).

1.1 Main Objective

The primary objective of this study is to examine the influence of corrections made to the FLORES dataset on the accuracy and reliability of machine translation (MT) model evaluations, with a particular emphasis on three low-resource African languages: Zulu, Hausa, and Northern Sotho. The contributions of this research are outlined as follows:

Contributions:

- **Analysis of Evaluation Metrics:** This study provides analysis of key evaluation metrics (e.g., BLEU, COMET, chrF, BERTScore), comparing their effectiveness in assessing translation quality and exploring their sensitivity to dataset variations.
- **Impact of Dataset Quality:** A comparative analysis of model performances using various versions of the FLORES dataset is presented,

illustrating how dataset quality influences MT outcomes and highlighting the importance of dataset correction for reliable evaluation.

- **Cross-Model Sensitivity Analysis:** The study investigates the sensitivity of different translation models to the various versions of the dataset, providing insight into how each model type reacts to changes in dataset quality and offering valuable information for model selection in low-resource settings.

- **Guidelines for Future Research:** Drawing from prior studies, we offer recommendations aimed at reducing errors and enhancing dataset quality in future machine translation benchmarks, contributing to the advancement of MT research, particularly for underrepresented languages.

1.2 Societal Value and Ethical Motivation

This study seeks to underscore the critical importance of ensuring the correctness and reliability of evaluation benchmarks, particularly those employed in machine translation. The integrity of such benchmarks directly influences the validity of model assessment and comparison; inaccuracies within these datasets may result in misleading conclusions regarding system performance. By concentrating on low-resource African languages, this work also foregrounds the ethical imperative to promote linguistic equity in computational research. Ensuring that all languages are treated with equal rigor and respect is essential for fostering inclusive and responsible development in Natural Language Processing (NLP).

2 Literature Survey

2.1 Prior work

Recent advances in machine translation evaluation have been driven by the need for comprehensive multilingual benchmarks, with FLORES-101 emerging as a foundational dataset in this space. Developed to address the scarcity of reliable evaluation data for low-resource languages, FLORES-101 (Goyal et al., 2022b) established a new standard by providing professionally translated parallel texts across 101 languages. Unlike previous benchmarks that focused primarily on high-resource language pairs, FLORES-101 systematically included languages from under-represented regions while maintaining strict quality control through native speaker verification(Goyal et al., 2022b).

Building on FLORES-101, the FLORES-200 dataset (Costa-Jussà et al., 2022) doubled the language coverage, incorporating 200 languages, including many critically under-resourced ones. This expansion aimed to reduce biases in MT evaluation and support research in long-tail languages. However, subsequent analyses revealed inconsistencies in translation quality, particularly for African languages, where cultural nuances and dialectal diversity introduced errors (Abdulmumin et al., 2024a).

To address these issues, the FLORES Fix for Africa initiative (Abdulmumin et al., 2024a) systematically refined translations for four African languages in FLORES-101/200. By involving native speakers and linguists, the project corrected lexical, syntactic, and cultural inaccuracies, yielding a more reliable benchmark. This effort highlighted the importance of human-in-the-loop validation for low-resource language data, a theme echoed in related work like MasakhaNEWS (Adelani et al., 2023).

2.2 Gap in pervious studies

Despite these advances, the field lacks rigorous studies on how dataset corrections—such as those in FLORES Fix—impact downstream model evaluation. For instance, do "fixed" datasets alter conclusions about model performance? This gap motivates our analysis of three African languages (Zulu, Hausa, Northern Sotho) to quantify the effect of data quality on evaluation metrics like BLEU and chrF++.

3 Methodology

3.1 Datasets

We evaluated translation performance using four datasets, each representing a different stage of refinement for low-resource African languages:

- **FLORES-101**: A standard benchmark for multilingual translation with aligned sentences in 101 languages.
  Available via Hugging Face. (Goyal et al., 2022a).

- **FLORES-200**: An extended version with broader language coverage.
  Available via GitHub (NLLB Team, 2022) (Goyal et al., 2021) (Guzmán et al., 2019).

- **FLORES Fix for Africa**: Community-corrected datasets for four African languages (Zulu, Hausa, Xitsonga, and Northern Sotho). Available via GitHub (Abdulmumin et al., 2024a) (Abdulmumin et al., 2024b)

- **FLORES+**: An independently corrected dataset with a focus on semantic and syntactic accuracy.
  Available via Hugging Face (NLLB Team et al., 2024).

Each dataset was structured into parallel sentence pairs consisting of an English source sentence and a corresponding African language reference translation.

3.2 Translation Models

We tested a range of multilingual and language-pair-specific translation models:

- **M2M100-418M**: A multilingual model trained for many-to-many translation without relying on English as the middle language.

- **NLLB-200-distilled-600M** and **NLLB-200-distilled-1.3B**: Models from Meta's No Language Left Behind initiative, optimized for low-resource language coverage via knowledge distillation.

- **OPUS-MT-en-ha** and **OPUS-MT-en-nso**: Direct English-to-Hausa and English-to-Northern Sotho translation models from the OPUS project.

All models were accessed via the Hugging Face Transformers library and loaded with corresponding tokenizers.

3.3 Preprocessing and Pipeline

The translation and evaluation pipeline was implemented as follows:

- **Dataset Loading**: Using the `datasets` library, all corpora were loaded into memory with splits for `dev` and `devtest`.

- **Dataset Structuring**: Source and reference texts were converted into dictionaries with `input_text` and `target_text` keys using utility functions such as `create_dataset_from_sentences`.

- **Translation Generation**: For each source sentence, translations were generated by passing tokenized inputs to the model, specifying the `forced_bos_token_id` for the target language.

- **Output Saving**: Generated translations were saved into files for each language–dataset–model combination for reproducibility and batch evaluation.

3.4 Evaluation Metrics

To evaluate translation quality, we used four automatic metrics:

- **BLEU** (Bilingual Evaluation Understudy): Evaluates n-gram overlap between model output and reference translations.

- **chrF**: Character-level F-score that better captures morphology and token structure in low-resource languages.

- **COMET**: A neural metric trained to predict human judgments of translation quality.

- **BERTScore**: Computes semantic similarity using contextual embeddings from BERT.

Each model was evaluated on each dataset variant across the three target languages (Zulu, Hausa, and Northern Sotho). The evaluation pipeline followed these steps:

1. Translate each English sentence in the `devtest` split using the selected model.

2. Compare the model output against the corrected and uncorrected reference translations using all four metrics.

3. Store and visualize the results for comparison, emphasizing the difference in performance due to dataset corrections.

An overview of the translation pipeline is shown in Appendix Figure 1.

4 Experiments and Results

The evaluation uses FLORES101, FLORES200, FLORES Fix for Africa, and FLORES+. However, it is important to note that FLORES101 and FLORES200 are essentially the same dataset, with FLORES200 containing additional languages. Similarly, FLORES Fix for Africa and FLORES+ are the same dataset, with the latter version offering more languages.

4.1 Table 1: M2M100-418M Dataset Comparison in Appendix Figure 2 and 3

This table compares the performance of the M2M100-418M model across four evaluation metrics—BLEU, chrF, BERTScore, and COMET—for three low-resource African languages: Hausa, Northern Sotho, and Zulu.

`Key Observations:`

- **Overall Performance Trends:** The performance across all languages and datasets is relatively low, which reflects the inherent challenges of handling low-resource languages in MT systems. The values for BLEU, chrF, and BERTScore indicate that the M2M100-418M model has limited success in accurately translating these languages.
- **Minimal Dataset Variation:** There is minimal variation in model performance across the different dataset versions for each language. For instance, the BLEU and chrF scores for FLORES101, FLORES200, FLORES Fix for Africa, and FLORES+ remain almost identical for most languages. This suggests that the improvements made through dataset corrections (in the FLORES Fix for Africa and FLORES+ versions) have not had a significant impact on the model's overall performance.
- **Impact of Dataset Refinements:** While the FLORES Fix for Africa and FLORES+ datasets show some minor improvements (notably for Zulu), these changes do not result in substantial gains across all evaluation metrics. For example, Zulu's BLEU score improved from 2.95 to 3.39, a notable increase, but other languages like Hausa and Northern Sotho saw little to no change.
- **COMET and BERTScore Stability:** The COMET and BERTScore metrics exhibit consistent values across all datasets for each language, indicating that the model's performance in terms of semantic similarity and human judgment prediction remains stable, regardless of dataset quality improvements.

Table 2: NLLB-200-distilled-600M Dataset Comparison in Appendix Figure 2 and Appendix Figure 3

This table evaluates the performance of the NLLB-200-distilled-600M model across the same metrics and dataset variants used for the previous analysis. The comparison is made for three low-resource African languages: Hausa, Northern Sotho, and Zulu.

- **Performance Consistency:** The NLLB-200-distilled-600M model shows stable performance across datasets, with minimal variation across BLEU, chrF, and BERTScore metrics.
- **Minor Decline in BLEU for Hausa and Zulu:** Hausa and Zulu show slight declines in BLEU with FLORES Fix for Africa and FLORES+ datasets, while Northern Sotho sees minor improvements.
- **Stable chrF and BERTScore:** Both metrics remain stable across datasets, indicating that dataset refinements did not affect translation quality.
- **COMET Consistency:** COMET scores remain consistent across datasets, suggesting no significant change in the model's predicted human judgment.

Table 3: NLLB-200-distilled-1.3B Dataset Comparison in Appendix Figure 3 and Appendix Figure 4

This table presents the performance results for the larger NLLB-200-distilled-1.3B model, which contains 1.3 billion parameters. The comparison is made for three low-resource African languages: Hausa, Northern Sotho, and Zulu.

- **Increased Performance with Larger Model:** The NLLB-200-distilled-1.3B model outperforms the smaller model in all evaluation metrics.
- **Minimal Impact of Dataset Variants:** Dataset refinements have minimal impact on the performance for all languages.
- **Minor Gains in Zulu and Northern Sotho:** Small improvements are observed in Zulu and Northern Sotho with the FLORES Fix for Africa and FLORES+ datasets.
- **COMET Stability:** COMET scores remain stable, indicating that the human-like judgment predicted by the model does not change across datasets.

This table presents the performance results for the larger NLLB-200-distilled-1.3B model, which contains 1.3 billion parameters. The table compares the model's performance across all dataset variants for three low-resource African languages: Hausa, Northern Sotho, and Zulu.

`Key Observations:`

- **Increased Performance with Larger Model:** The NLLB-200-distilled-1.3B model shows significantly higher performance across all metrics (BLEU, chrF, BERTScore, COMET) compared to the smaller NLLB-200-distilled-600M model. For instance, Hausa shows a BLEU score of 25.24 with FLORES101 and FLORES200, which is a noticeable improvement over the smaller model's performance.
- **Minimal Impact of Dataset Variants:** Similar to the previous evaluations, the dataset variants (FLORES101, FLORES200, FLORES Fix for Africa, and FLORES+) show minimal variation in performance. There is no significant improvement or decline in scores across the datasets, suggesting that the fixes in the FLORES Fix for Africa and FLORES+ datasets had little effect on the overall performance of the model.
- **Minor Gains in Zulu and Northern Sotho:** The FLORES Fix for Africa and FLORES+ datasets do show some minor improvements in Zulu and Northern Sotho scores, particularly in chrF and BERTScore. For Zulu, chrF increased from 58.12 to 59.14, and BERTScore remained stable at 0.84 across datasets, showing slight enhancement in translation quality with the corrected datasets.
- **COMET Stability:** The COMET metric, which predicts human judgments, remains stable across all dataset variants. This suggests that the refinements in dataset translations did not significantly alter the model's alignment with human judgment.

.

Table 4: OPUS-MT-en-ha Dataset Comparison (Hausa) in Appendix Figure 4 and Appendix Figure 5

This table shows evaluation results for the English-to-Hausa translation model (OPUS-MT).

- **Stable Performance Across Datasets:** The OPUS-MT-en-ha model shows moderate and consistent performance across all datasets. Minimal variation is observed between FLORES101, FLORES200, FLORES Fix for Africa, and FLORES+ datasets.
- **Minor chrF Decline in Corrected Datasets:** There is a small drop in chrF for FLORES Fix for Africa and FLORES+, but overall performance remains stable.
- **Stable BERTScore and COMET:** Both BERTScore and COMET metrics exhibit consistent values, indicating no significant impact from dataset refinements on the model's predicted semantic accuracy and human judgment.

Table 5: OPUS-MT-en-nso Dataset Comparison (Northern Sotho) in Appendix Figure 5 and Appendix Figure 6

This table presents evaluation results for the English-to-Northern Sotho translation model (OPUS-MT).

- **Stable Performance Across Datasets:** The OPUS-MT-en-nso model demonstrates consistent performance across all datasets. FLORES101 and FLORES200 show stable BLEU and chrF scores.
- **Minor Gains in BLEU for Northern Sotho:** A slight improvement in BLEU and chrF is observed with the FLORES Fix for Africa and FLORES+ datasets.
- **Stable BERTScore and COMET:** As with the Hausa model, BERTScore and COMET remain stable, indicating no major effect from dataset corrections.

5 Reflections and Discussion

The evaluation of various models across different datasets has led to several key reflections on the role of model capacity, dataset quality, and language-specific factors in machine translation (MT) performance.

- **Model Capacity and Performance:** As expected, model size significantly influences translation quality. The NLLB-200-distilled-1.3B model, with its 1.3 billion parameters,

consistently outperformed the smaller NLLB-200-distilled-600M model across all metrics (BLEU, chrF, BERTScore, and COMET). The larger model's ability to capture more nuanced language features and better handle low-resource languages was evident in the improved scores, particularly for Hausa and Zulu.

- **Dataset Quality and Impact:** The refinement of datasets—especially the FLORES Fix for Africa and FLORES+ versions—demonstrated the importance of dataset quality in MT model evaluations. While the impact was not universally transformative, there were notable improvements, particularly for Zulu and Northern Sotho. These improvements were observed in BLEU and chrF scores, reflecting enhanced translation accuracy. This finding emphasises the need for high-quality, corrected datasets to achieve more reliable MT evaluations for under-represented languages.

- **Language Sensitivity to Dataset Corrections:** Dataset refinements proved to be language-sensitive. Zulu exhibited the most significant improvements in translation quality, while Hausa and Northern Sotho showed more minimal changes. This highlights the challenges of applying dataset corrections universally and underscores the need for tailored strategies when working with different languages.

- **Model Sensitivity to Dataset Corrections:** While larger models demonstrated better performance, their sensitivity to dataset corrections varied. In particular, the OPUS-MT models, despite showing stable performance across datasets, were less responsive to dataset refinements. The slight variation in their evaluation metrics, particularly for Hausa and Northern Sotho, suggests that these models are relatively robust but less influenced by the quality of dataset corrections.

6 Conclusion

In conclusion, several key insights emerged from our evaluation. First, model capacity plays a crucial role—larger models with more parameters (such as the 1.3 billion parameter model) consistently outperformed their smaller counterparts. Second, dataset quality matters. The corrected datasets—FLORES Fix for Africa and FLORES+—often resulted in improved BLEU and chrF scores, particularly for Zulu and Northern Sotho. Third, the impact of dataset corrections is language-sensitive, with Zulu showing the most notable gains. Lastly, we observed model sensitivity differences: while OPUS-MT models are generally robust, they demonstrated minimal responsiveness to the corrected datasets, showing little to no change in evaluation metrics.

6.1 Future Work

The findings of this study pave the way for several avenues of future research and improvements in machine translation (MT) systems for low-resource languages:

- **Adding More Languages:** Future work could focus on expanding the number of low-resource languages included in the dataset. By incorporating additional languages, particularly from underrepresented regions, MT systems can be made more inclusive and versatile.

- **Additional Preprocessing and Model Training:** Further preprocessing steps, such as domain-specific fine-tuning and training models on larger, more diverse datasets, could improve the performance of MT systems. Training models specifically tailored to different language families or domains could lead to more accurate translations for specialized contexts.

- **Region-Specific Dataset Refinements:** As demonstrated in this study, dataset corrections are language-specific. Future work could include region-specific datasets, where refinements are made to better address regional linguistic variations and nuances, further enhancing translation accuracy.

- **Post-Processing of Translations:** Post-processing techniques, such as grammar correction, fluency enhancement, and style adaptation, could be implemented to refine machine-generated translations. This would help ensure that translations are not only accurate but also fluent and culturally appropriate for the target audience.

.

In summary, this study underscores the crucial role of both model capacity and dataset quality in determining the performance of MT systems. It also highlights the need for continued focus on dataset corrections and the tailoring of MT models to specific languages, especially low-resource ones.

7 Github

Code is Available via GitHub

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024a. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse S. Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024b. Correcting flores evaluation dataset for four african languages. *Preprint*, arXiv:2409.00626.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, and 1 others. 2023. Masakhanews: News topic classification for african languages. *arXiv preprint arXiv:2304.09972*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022a. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022b. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english.

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, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

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 NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

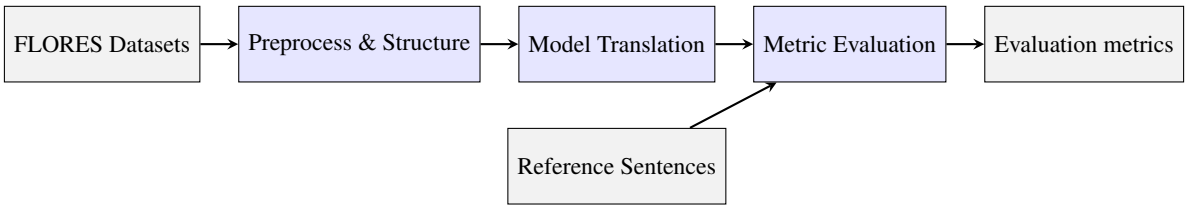

Figure 1: Translation Evaluation Pipeline

Table 1: Dataset Comparison for m2m100_418M across Evaluation Metrics

Language	Dataset	BLEU	chrF	BERTScore	COMET
Hausa	FLORES101	4.65	26.24	0.71	0.03
	FLORES200	4.65	26.24	0.71	0.03
	FLORES Fix for Africa	4.62	26.24	0.71	0.03
	FLORES+	4.62	26.24	0.71	0.03
Northern Sotho	FLORES101	3.34	24.12	0.73	-0.08
	FLORES200	3.34	24.12	0.73	-0.08
	FLORES Fix for Africa	3.31	24.13	0.73	-0.08
	FLORES+	3.31	24.13	0.73	-0.08
Zulu	FLORES101	2.95	26.97	0.70	0.06
	FLORES200	2.95	26.97	0.70	0.06
	FLORES Fix for Africa	3.39	27.44	0.71	0.06
	FLORES+	3.39	27.44	0.71	0.06

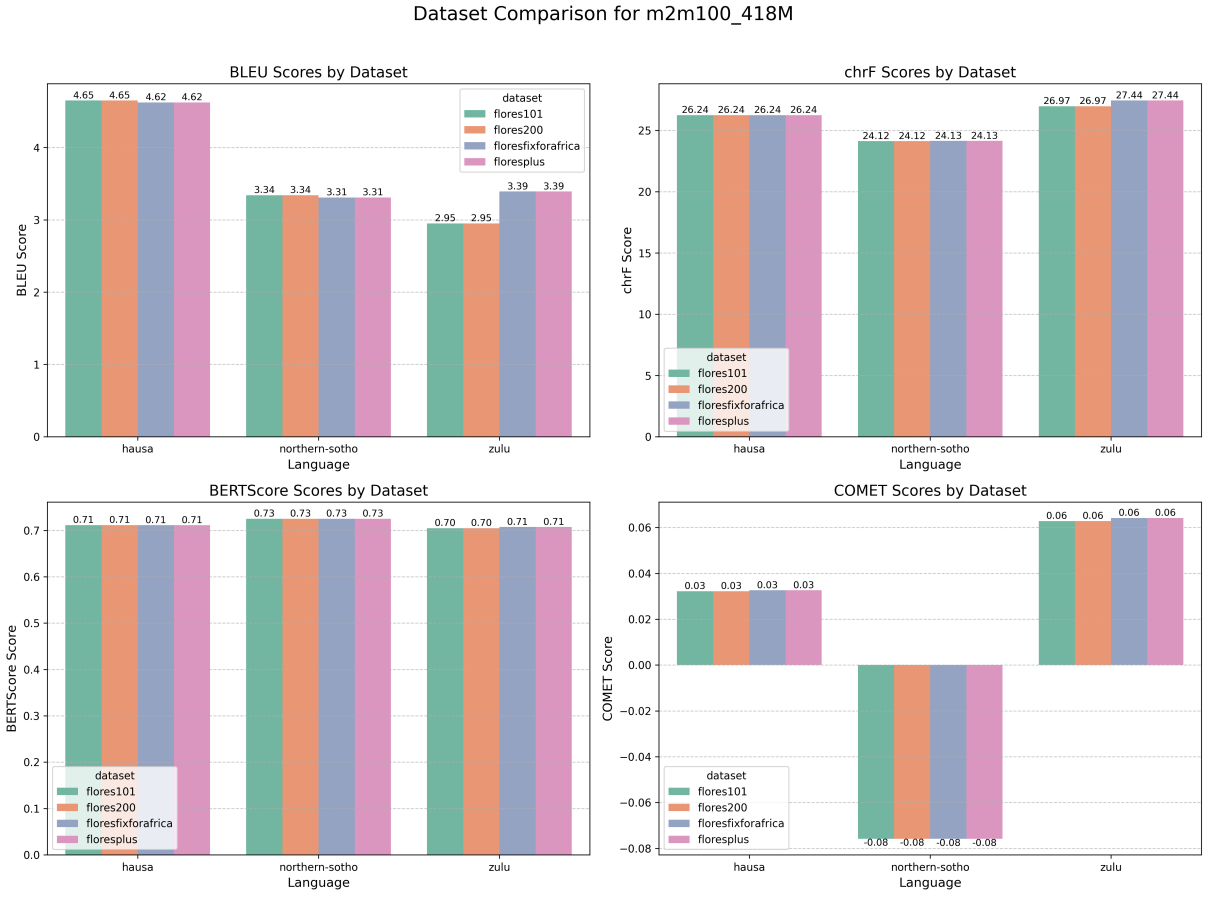

Figure 2: M2M100-418M Dataset Comparison Bar-charts

Table 2: Dataset Comparison for nllb-200-distilled-600M across Evaluation Metrics

Language	Dataset	BLEU	chrF	BERTScore	COMET
Hausa	FLORES101	23.80	51.36	0.82	0.64
	FLORES200	23.80	51.36	0.82	0.64
	FLORES Fix for Africa	23.45	51.08	0.82	0.64
	FLORES+	23.45	51.08	0.82	0.64
Northern Sotho	FLORES101	21.75	51.05	0.81	0.67
	FLORES200	21.75	51.05	0.81	0.67
	FLORES Fix for Africa	22.06	51.28	0.81	0.67
	FLORES+	22.06	51.28	0.81	0.67
Zulu	FLORES101	16.82	56.27	0.83	0.67
	FLORES200	16.82	56.27	0.83	0.67
	FLORES Fix for Africa	17.83	57.14	0.84	0.67
	FLORES+	17.83	57.14	0.84	0.67

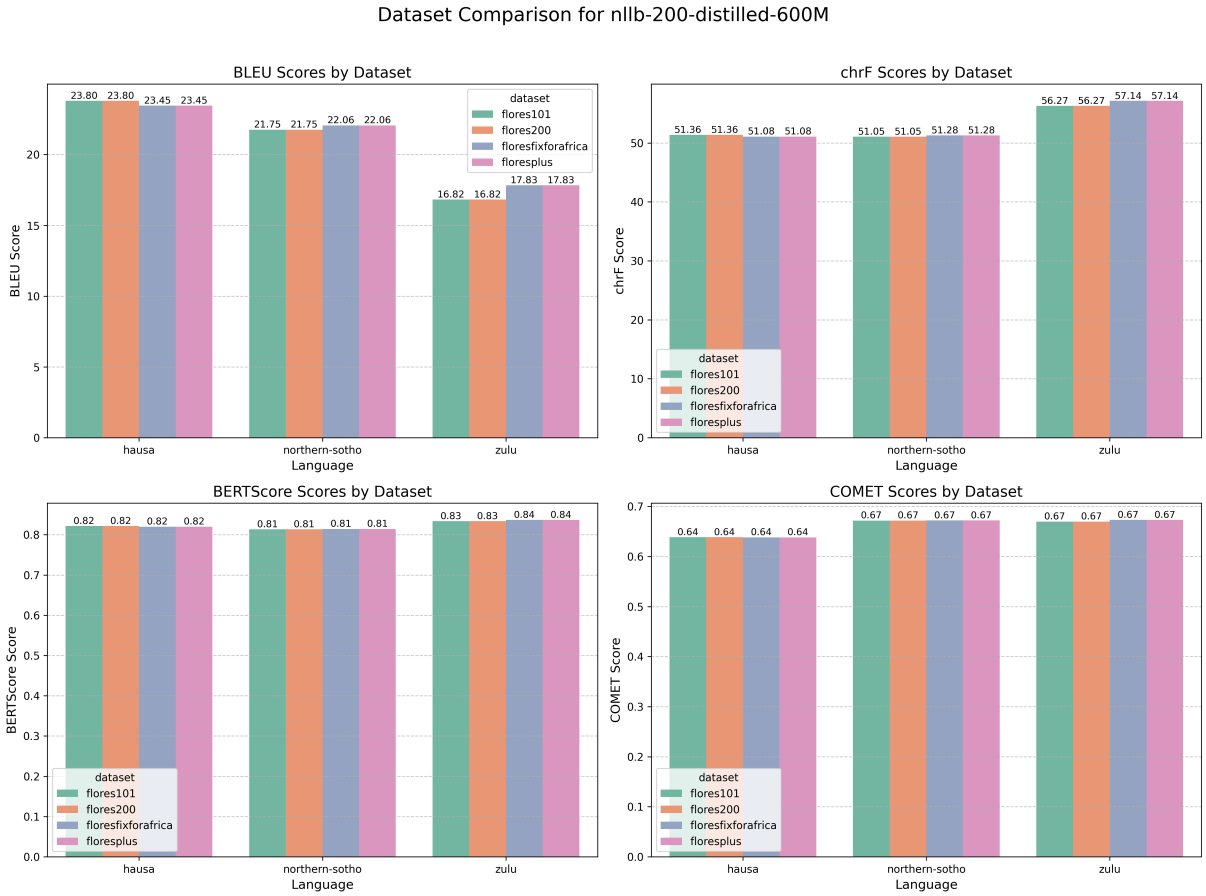

Figure 3: nllb-200-distilled-600M Dataset Comparison Bar-charts

Table 3: Dataset Comparison for nllb-200-distilled-1.3B across Evaluation Metrics

Language	Dataset	BLEU	chrF	BERTScore	COMET
Hausa	FLORES101	25.24	53.04	0.83	0.65
	FLORES200	25.24	53.04	0.83	0.65
	FLORES Fix for Africa	24.74	52.65	0.83	0.65
	FLORES+	24.74	52.65	0.83	0.65
Northern Sotho	FLORES101	22.69	51.96	0.82	0.68
	FLORES200	22.69	51.96	0.82	0.68
	FLORES Fix for Africa	22.97	52.15	0.82	0.68
	FLORES+	22.97	52.15	0.82	0.68
Zulu	FLORES101	18.47	58.12	0.84	0.69
	FLORES200	18.47	58.12	0.84	0.69
	FLORES Fix for Africa	18.57	59.14	0.84	0.69
	FLORES+	18.57	59.14	0.84	0.69

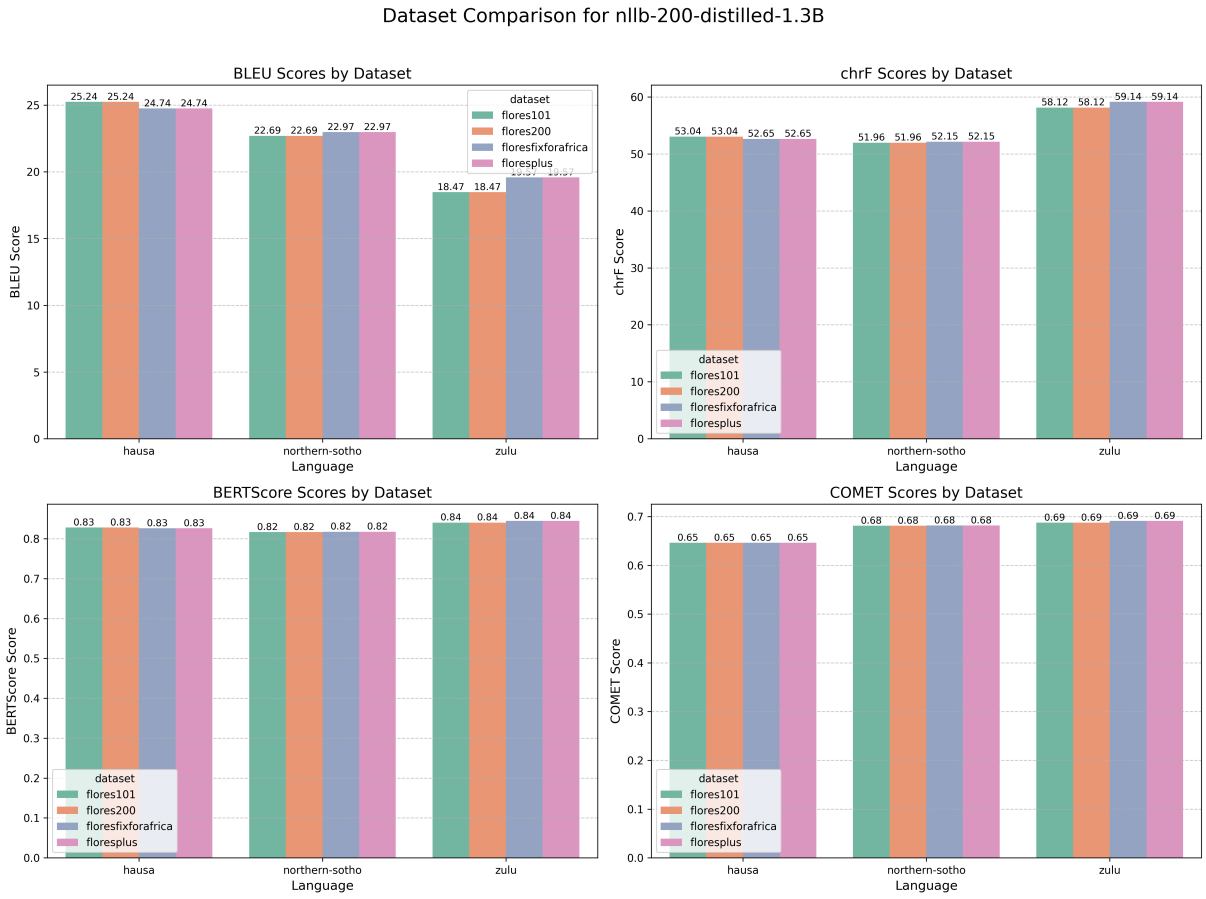

Figure 4: nllb-200-distilled-1.3BM Dataset Comparison Bar-charts

Table 4: Dataset Comparison for `opus-mt-en-ha` on Hausa

Dataset	BLEU	chrF	BERTScore	COMET
FLORES101	7.23	31.54	0.75	0.07
FLORES200	7.23	31.54	0.75	0.07
FLORES Fix for Africa	7.23	31.47	0.75	0.07
FLORES+	7.23	31.47	0.75	0.07

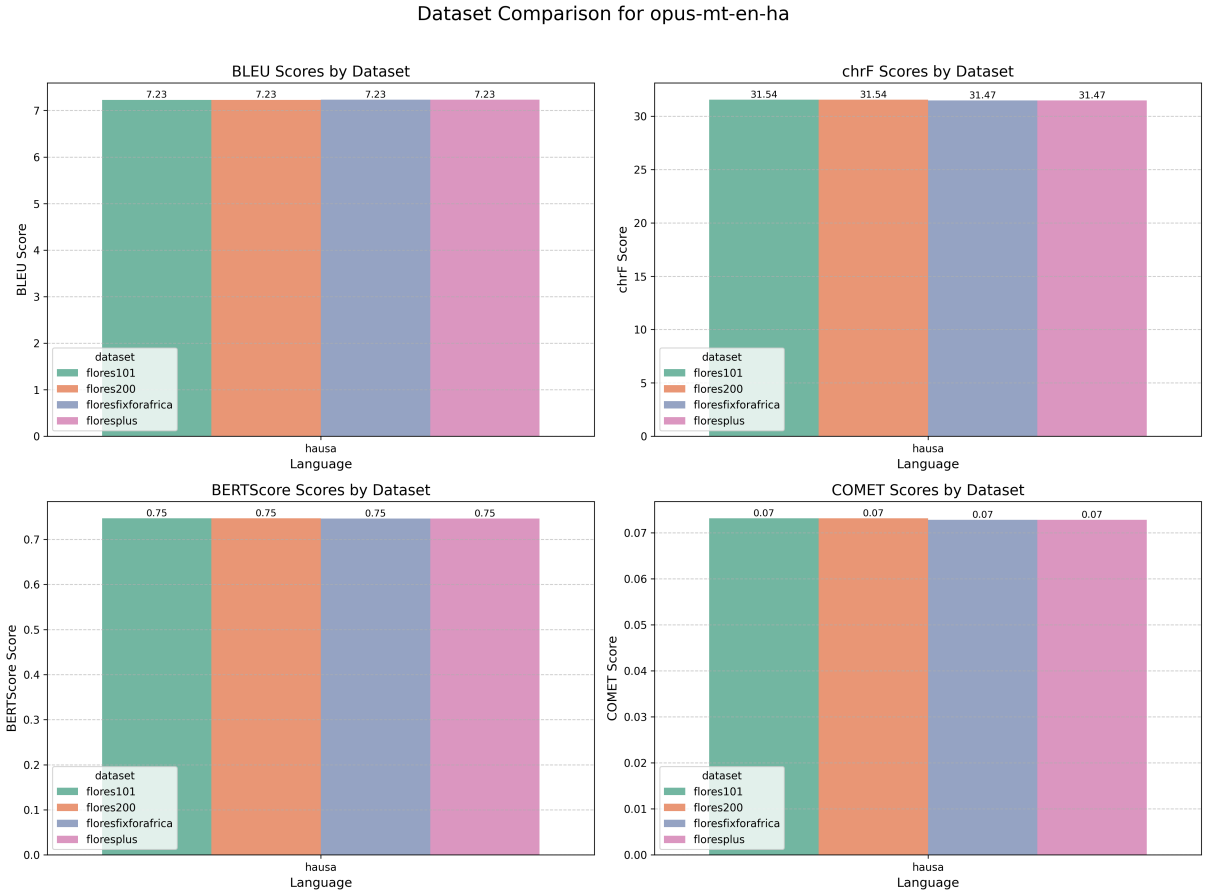

Figure 5: opus-mt-en-ha Dataset Comparison Bar-charts

Table 5: Dataset Comparison for opus-mt-en-nso on Northern Sotho

Dataset	BLEU	chrF	BERTScore	COMET
FLORES101	15.97	43.95	0.78	0.41
FLORES200	15.97	43.95	0.78	0.41
FLORES Fix for Africa	16.11	44.08	0.78	0.41
FLORES+	16.11	44.08	0.78	0.41

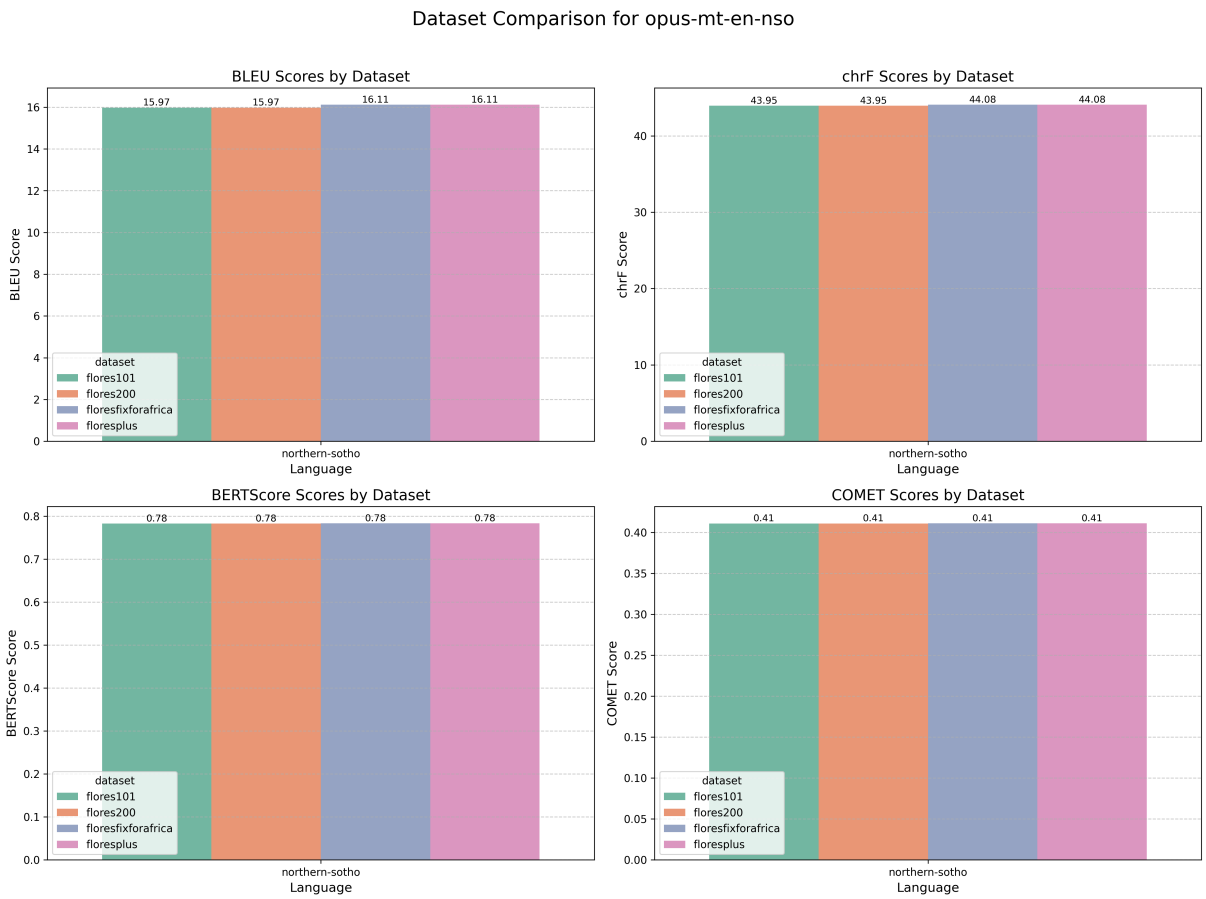

Figure 6: opus-mt-en-nso Dataset Comparison Bar-charts