Evaluating WMT 2025 Metrics Shared Task Submissions on the SSA-MTE African Challenge Set

Senyu Li[1,2] **Felermino D. M. A. Ali**[6,7] **Jiayi Wang**[3]
Rui Sousa-Silva[7] **Henrique Lopes Cardoso**[6] **Pontus Stenetorp**[3,5]
Colin Cherry[8] **David Ifeoluwa Adelani**[1,2,4]
[1]Mila - Quebec AI Institute, [2]McGill University, [3]University College London,
[4]Canada CIFAR AI Chair, [5]LLMC, National Institute of Informatics
[6]LIACC, Faculdade de Engenharia, Universidade do Porto
[7]CLUP, Faculdade de Letras, Universidade do Porto [8]Google
{senyu.li, david.adelani}@mila.quebec
jiaywang@cs.ucl.ac.uk up202100778@fe.up.pt

Abstract

This paper presents the evaluation of submissions to the WMT 2025 Metrics Shared Task on the SSA-MTE challenge set, a large-scale benchmark for machine translation evaluation (MTE) in Sub-Saharan African languages. The SSA-MTE test sets contains over 12,768 human-annotated adequacy scores across 11 language pairs sourced from English, French, and Portuguese, spanning six commercial and open-source MT systems. Results show that correlations with human judgments remain generally low, with most systems falling below the 0.4 Spearman threshold for medium-level agreement. Performance varies widely across language pairs, with most correlations under 0.4; in some extremely low-resource cases, such as Portuguese–Emakhuwa, correlations drop to around 0.1, underscoring the difficulty of evaluating MT for very low-resource African languages. These findings highlight the need for more robust and generalizable evaluation methods tailored to African language contexts.

1 Introduction

In recent years, with the rise of large language models (LLMs), more and more machine translation (MT) systems have emerged, demonstrating competitive performance. This growth has created an increasingly urgent need for more accurate methods to assess the quality of generated translations. Traditional metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and chrF (Popović, 2015), which rely on n-gram matching, show only limited correlation with human judgments, indicating their limited ability to capture semantic-level quality (Callison-Burch et al., 2006).

More recently, neural metrics such as BERTScore (Zhang et al., 2020a) have shown improved capability in capturing semantic similarity (Freitag et al., 2024; Zhang et al., 2020b). COMET (Rei et al., 2020) further advances this by framing machine translation evaluation (MTE) as a regression task using encoder-only language models and training on human-annotated scores. Likewise, the MetricX (Juraska et al., 2023) family of metric, based on the mT5 (Xue et al., 2020) series multilingual encoder-decoder LM, adopts a regression-based framework similar to COMET. These neural, learned metrics have been shown to achieve higher correlations with human assessments across a wide range of languages (Rei et al., 2020; Juraska et al., 2023).

However, before 2024, due to the lack of machine translation evaluation data, the performance of these models was largely untested on Sub-Saharan African languages. In 2024, AfriMTE (Wang et al., 2024a) was created. Evaluation results revealed that while existing metrics performed well on some relatively higher-resourced languages, they struggled with translations for very low-resource languages such as Twi and Luo (Wang et al., 2024b). The authors demonstrated that this gap can be partially addressed by further pretraining models on African languages; however, performance remains low for specific language pairs, like eng-luo, underscoring the need for dedicated training data for these languages.

Although AfriMTE marked progress in this area, several limitations remained. The dataset was relatively small, with only about 200 cases per language pair. Also, AfriMTE included outputs only from NLLB-200 (600M) (NLLB-Team et al., 2022) and M2M-100 (418M) (Fan et al., 2021). In addition, AfriMTE did not provide any training data, which meant that neural-based metrics could not be directly optimized for Sub-Saharan languages.

To address these challenges, Li et al. (2025) in-

Proceedings of the Tenth Conference on Machine Translation, pages 913–919

troduced SSA-MTE, a larger-scale dataset created following the same protocol and using the same annotation tool as AfriMTE. SSA-MTE covers 13 Sub-Saharan African languages, with test sets for 10 languages and training sets for 12. It features several source languages—English, French, and Portuguese—representing the Anglophone, Francophone, and Lusophone linguistic communities in the region. SSA-MTE includes translations from six MT systems and contains over 73,000 annotations in total.

The WMT 2025 Metrics Shared Task incorporates the SSA-MTE test set as a dedicated challenge set, enabling the evaluation of MT metrics on low-resource African languages. This inclusion establishes a benchmark for assessing the ability of MTE systems to generalize across under-resourced African languages.

2 SSA-MTE

We perform our evaluation on the recently released **SSA-MTE** dataset (Li et al., 2025), a large-scale human-annotated benchmark for assessing machine translation quality for African languages. The dataset contains over 73,000 sentence-level annotations across 13 language pairs (LPs) in the news domain, of which 10 LPs include a dedicated test set.

The test set covers 7 English–sourced LPs: Amharic (`eng-amh`), Hausa (`eng-hau`), Kikuyu (`eng-kik`), Kinyarwanda (`eng-kin`), Luo (`eng-luo`), Twi (`eng-twi`), and Yorùbá (`eng-yor`); 2 French–sourced LPs: Ewe (`fra-ewe`) and Wolof (`fra-wol`); and 1 Portuguese–sourced LP: Emakhuwa (`por-vmw`). For this challenge set, the authors additionally introduced another Portuguese–sourced LP: Nyanja (`por-nya`). The size of each test set is shown in Table 1.

The selected LPs reflect Africa's linguistic and regional diversity, covering Anglophone, Francophone, and Lusophone areas. In addition, the languages span three major language families: Afro-Asiatic (Hausa, Amharic), Niger–Congo (Kikuyu, Kinyarwanda, Emakhuwa, Nyanja, Twi, Yorùbá, Ewe, Wolof), and Nilo-saharan (Luo), ensuring representation across West, East, Central, and Southern Africa, and thus capturing both geographic and typological diversity.

Each instance is annotated by one human evaluator with both a continuous adequacy score and span-level (character-based) error labels, enabling fine-grained evaluation of MT outputs. Annotators are provided with the source sentence and its translation; instructed to first identify all errors according to the annotation protocol proposed by Wang et al., and then assign a final adequacy score.

For English- and French-sourced cases, source sentences are drawn from the Global Voices news website, following harmful-content filtering and topic-diversity-based article selection (Li et al., 2025). For Portuguese-sourced cases, source sentences are extracted from the Multilingual Open Text dataset, which features news articles published by Voice of America (VOA). All source texts are translated into the target languages by professional translators to serve as reference translations (Ali et al., 2024).

For English- and French-sourced sentences, six MT systems are included (but only five for Kikuyu, as Google Translate does not support it): GPT-4o, Gemini-1.5, Claude-3.5, Google Translate, and two open-source models: NLLB-200-distilled-600M (NLLB-Team et al., 2022) and M2M-100-418M (Fan et al., 2021). Each system contributes an equal number of translations. For Portuguese-sourced sentences, four MT systems are included: GPT-4o-mini, Gemini-1.5, Claude-3.5, and Google Translate.

Compared to the African Challenge Set in the WMT 2024 Metrics Shared Task, this year's set is substantially larger (12,768 vs. 2,815 annotated instances), covers a broader range of MT systems, and introduces Portuguese as a new source language. This broader linguistic and system coverage is expected to yield a more accurate and comprehensive evaluation of metric performance for African languages.

3 Metrics

The submissions of WMT 2025 Metrics Shared Task contain baseline metric results provided by the organizers, as well as the results of primary and secondary submissions from participants' metric systems. This section introduces each of these approaches[1].

3.1 Baselines

We received the following baseline metrics from the organizers: BERTScore (Zhang et al.,

[1]Detailed information of the submissions has not yet been provided by the organizer, and will be added in the camera-ready version.

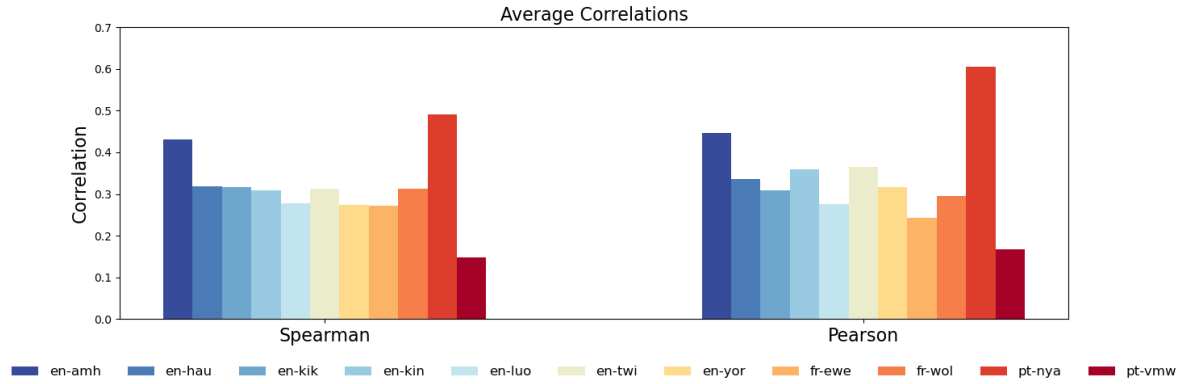

Figure 1: **Average Spearman and Pearson correlations across LPs**. AfriCOMET-1.1-MTL and SSA-COMET-MTL were not included in the calculation.

Language Pair	Size
en-amh	1,166
en-hau	1,192
en-kik	1,172
en-kin	1,210
en-luo	1,199
en-twi	1,200
en-yor	1,206
fr-ewe	1,077
fr-wol	1,175
pt-nya	1,241
pt-vmw	930
Total	12,768

Table 1: Number of instances per language pair.

2020b), BLEU (Papineni et al., 2002), chrF (Popović, 2015), COMET22 (Rei et al., 2022a), COMETKiwi22 (Rei et al., 2022b), Sentinel-Cand (Perrella et al., 2024), Sentinel-Src (Proietti et al., 2025), spBLEU (Fan et al., 2020), and YiSi-1 (Lo, 2019).

3.2 Submissions

From each participating team, we received one primary submission and several secondary submissions. The primary submissions included MetricX-25 (Juraska et al., 2025), mr7.2.1 (Hrabal et al., 2025), Polycand-2 (Züfle et al., 2025), and rankedCOMET (Maharjan and Shrestha, 2025). The secondary submissions included baseCOMET, MetricX-25-QE, MetricX-25-Ref, Polycand-1, and Polyc-3.

MetricX-25 MetricX-25 is an encoder-only regression model initialized from Gemma3 12B (Team et al., 2025) and fine-tuned on WMT15–23 DA and MQM scores in two stages: first on z-normalized DA scores, then on a 1:1 mixture of rescaled DA and MQM scores with a score-type indicator to align outputs with the intended evaluation (ESA/DA vs. MQM). Compared to MetricX-24 (Juraska et al., 2024), this setup increases the weight of DA data in the second stage and explicitly conditions on score type. Both stages also include a small proportion of synthetic WMT-derived data used in MetricX-24. Training uses a maximum input length of 4K tokens to balance performance and coverage of low-resource language examples.

mr7.2.1 uses Gemma3 27B (Team et al., 2025), prompted with the DSPy framework and optimized with MIPROv2 (Opsahl-Ong et al., 2024), first generating seven aspect scores (0–10) and then producing the final overall score (0–100).

rankedCOMET is based on COMET22 (Rei et al., 2022a) used in zero-shot inference, producing raw segment-level scores that are then adjusted with per-language-pair rank normalization, yielding calibrated distributions and improved correlation with evaluation metrics.

Polycand-2 is a COMET-poly supervised model trained on WMT data up to 2024 (DA/ESA/MQM merged on a single scale). It extends COMET by using two alternative translations of the same source to provide better context for scoring. English–Korean and Japanese–Chinese were excluded from training.

3.3 AfriCOMET-1.1-MTL and SSA-COMET-MTL

For the WMT 2024 Metrics Shared Task on the African Challenge Set, the authors explored replacing the original AfroXLMR (Alabi et al., 2022) with an enhanced African-centric multilingual pretrained encoder, AfroXLMR-76L (Adelani et al.,

2024)[2], to build MT evaluation and QE models tailored for African languages, and named the resulting models as AfriCOMET-1.1 (Wang et al., 2024b). This upgrade yields notable improvements: AfriCOMET achieved the highest Spearman correlation on the 2024 WMT African Challenge Set among all benchmarked systems (unweighted setting), underscoring the critical role of a stronger base encoder in advancing the quality of MTE for African languages.

Since this challenge set was built on SSA-MTE, which provides a large-scale training set, we further explored the impact of incorporating in-domain, in-task training data on final model performance. Specifically, we report results for SSA-COMET-MTL (MTL stands for multi-task learning), currently the best SSA-COMET model, built using the same pipeline as AfriCOMET-1.1 with the same base model and training data, but augmented with additional training examples from SSA-MTE. As the original authors did not provide an MTL version of AfriCOMET-1.1, we reproduced it by following the same pipeline and hyperparameters, enabling a fair comparison. We chose to compare using the MTL versions of the models because, empirically, MTL models tend to outperform STL models when all other factors are held constant (Li et al., 2025).

4 Analysis

Table 2 shows the average segment-level correlations on the SSA-MTE challenge set. Overall, correlations are moderate, with most Spearman values between 0.3 and 0.5 and Pearson values between 0.35 and 0.55. We adapt the following definition of levels of agreement: a Spearman and a Pearson correlation lower than 0.4 indicates a low-level agreement, a value between 0.4 and 0.6 indicates medium-level agreement, and a value greater than 0.6 indicates high-level agreement with human judgments.

4.1 Official Baselines

Among the official baselines provided by the organizers, chrF and YiSi-1 achieve the strongest overall performance (Spearman 0.506 / 0.460, Pearson 0.532 / 0.493), indicating that carefully tuned lexical and embedding similarity measures remain competitive in this evaluation setting. Notably, chrF even has comparable performance to the best supervised participant submission, MetricX-25. spBLEU shows a notable advantage over BLEU in Spearman (0.434 vs. 0.389), suggesting better ranking stability when using sentencepiece tokenization for morphologically rich or orthographically diverse languages. COMET22 achieves a lower correlation (0.363 in Spearman) compared to AfriCOMET-1.1-MTL, which shares the same architecture but uses an African-language–enhanced encoder LM (AfroXLMR-76L). This mirrors last year's findings on the importance of the base model. COMETKiwi22, a reference-free system, exhibits a clear performance drop compared to its reference-present variant COMET22, indicating that reference information remains important for neural, learned metrics. The Sentinel metrics score lowest overall, suggesting that their coarse-grained features are insufficient for fine-grained adequacy judgments in this domain.

Metrics	Pearson	Spearman
Baseline		
chrF	0.532	0.506
YiSi-1	0.493	0.460
spBLEU	0.395	0.434
BERTScore	0.471	0.425
BLEU	0.336	0.389
COMET22	0.405	0.363
COMETKiwi22	0.253	0.244
sentinel-cand	0.107	0.102
sentinel-src	0.068	0.073
Primary		
MetricX-25	0.530	0.467
mr7.2.1	0.477	0.380
rankedCOMET	0.377	0.364
Polycand-2	0.142	0.132
Secondary		
MetricX-25-Ref	0.550	0.490
MetricX-25-QE	0.490	0.427
baseCOMET	0.405	0.364
Polyic-3	0.177	0.144
Polycand-1	0.159	0.152
Additional		
SSA-COMET-MTL	0.688	0.630
AfriCOMET-1.1-MTL	0.599	0.552

Table 2: Average segment-level correlation coefficients of MT evaluation metrics across languages on the SSA-MTE test set.

4.2 Participant Submissions

For participant submissions, the highest average correlations come from MetricX-25-Ref (Pearson 0.550, Spearman 0.490), outperforming both its QE variant (MetricX-25-QE) and its default vari-

[2] https://huggingface.co/Davlan/afro-xlmr-large-76L

ant (MetricX-25). This reinforces the finding that reference-based approaches are more effective than QE-only approaches in the SSA-MTE setting. MetricX-25 is the best among primary submissions, followed closely by MetricX-25-QE and rankedCOMET. mr7.2.1 is competitive in Pearson but drops in Spearman, while the Polycand/Polyic series performs notably lower, suggesting limited adaptation to the SSA-MTE adequacy signal.

4.3 Additional Baselines

We also include two other baselines intended to benchmark progress in African-language MT evaluation: AfriCOMET-1.1-MTL and SSA-COMET-MTL. Both use the same base model and pipeline, with SSA-COMET-MTL further incorporating in-domain SSA-MTE training data. SSA-COMET-MTL achieves the highest overall scores in this evaluation (Pearson 0.688, Spearman 0.630), outperforming AfriCOMET-1.1-MTL by +0.089 Pearson and +0.078 Spearman. These gains highlight the value of in-domain, in-task supervision. Among submitted systems, MetricX-25 and its variants achieve the highest correlations, generally reaching a medium-level agreement with human judgments. All other submissions remain in the low-agreement range (Spearman < 0.4), indicating notable room for improvement.

4.4 Per-LP Performance

Figure 1 presents per-language averages across all metrics. Performance varies substantially by language pair: pt-nya is the easiest (Pearson 0.606, Spearman 0.491), while pt-vmw is the hardest (Pearson 0.168, Spearman 0.147). English-sourced pairs show mixed difficulty, with en-amh and en-twi at the higher end, and en-luo and en-yor lower. Among French-sourced pairs, fr-wol tends to outperform fr-ewe. These trends suggest that both source–target pairing and specific linguistic features of the target language influence evaluation difficulty. However, only en-amh and pt-nya achieve medium-level Spearman correlations; all other LPs remain in the low-agreement range, with pt-vmw extremely low (around 0.1 on average).

In summary, correlations on the SSA-MTE challenge set remain moderate even for the strongest systems, underscoring the difficulty of MT evaluation for low-resource African languages. The results indicate that reference-based learned metrics with in-domain training (e.g., SSA-COMET-MTL, MetricX-25-Ref) offer clear advantages, but there is significant room for improvement before reaching high-agreement levels with human judgment. Promising directions include the use of African-language–enhanced encoders and leveraging large-scale in-domain, in-task training data.

5 Conclusion

We have presented the results of the WMT 2025 Metrics Shared Task for the SSA-MTE challenge set, the largest and most diverse benchmark to date for MT evaluation in Sub-Saharan African languages. The evaluation covered a wide range of metrics, including official baselines, participant submissions, and additional African-focused baselines. Overall, correlations with human judgments remain modest, with most submitted systems achieving Spearman correlations below 0.4, indicating low-level agreement. Reference-based neural metrics generally outperform reference-free approaches, with MetricX-25-Ref leading among participant systems. SSA-COMET-MTL, trained with in-domain SSA-MTE data, sets a new reference point for African-language MTE, demonstrating clear gains over AfriCOMET-1.1-MTL and underscoring the value of domain-matched supervision. Per-language analysis shows substantial variation in difficulty, reflecting differences in source–target pairing, linguistic complexity, and resource availability. Only en-amh and pt-nya surpass the 0.4 threshold for medium-level agreement, while pt-vmw remains particularly challenging with correlations near 0.1. These findings suggest that while progress has been made, significant room remains to improve robustness and accuracy for the most difficult language pairs. Future work should explore integrating African-language–enhanced encoders, expanding the diversity of training data, and developing methods that can better generalize across very low-resource languages.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. 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.

Felermino D. M. A. Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. Building resources for emakhuwa: Machine translation and news classification benchmarks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14842–14857, Miami, Florida, USA. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. 22(1).

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Miroslav Hrabal, Ondrej Glembek, Aleš Tamchyna, Almut Silja Hildebrand, Alan Eckhard, Miroslav Štola, Sergio Penkale, Zuzana Šimečková, Ondřej Bojar, Alon Lavie, and Craig Stewart. 2025. Cuni and phrase at wmt25 mt evaluation task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. Metricx-25 and gemspaneval: Google translate submissions to the wmt25 evaluation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Senyu Li, Jiayi Wang, Felermino D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. Ssa-comet: Do llms outperform learned metrics in evaluating mt for under-resourced african languages? *Preprint*, arXiv:2506.04557.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Sujal Maharjan and Astha Shrestha. 2025. Ranked-comet: Elevating a 2022 baseline to a top-5 finish in the wmt 2025 qe task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

NLLB-Team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. *Preprint*, arXiv:2406.11695.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand. 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.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. *Preprint*, arXiv:2508.10175.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022a. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, and 39 others. 2024a. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenetorp. 2024b. Evaluating WMT 2024 metrics shared task submissions on AfriMTE (the African challenge set). In *Proceedings of the Ninth Conference on Machine Translation*, pages 505–516, Miami, Florida, USA. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues, and Mrinmaya Sachan. 2025. Comet-poly: Machine translation metric grounded in other candidates. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.