Evaluating WMT 2024 Metrics Shared Task Submissions on AfriMTE (the African Challenge Set)

Jiayi Wang[1], **David Ifeoluwa Adelani**[2,3,4], **Pontus Stenetorp**[1]
[1]University College London, UK, [2]Mila - Quebec AI Institute, Canada
[3]McGill University, Canada, [4]Canada CIFAR AI Chair
ucabj45@ucl.ac.uk, david.adelani@mila.quebec, p.stenetorp@cs.ucl.ac.uk

Abstract

The AFRIMTE challenge set from WMT 2024 Metrics Shared Task aims to evaluate the capabilities of evaluation metrics for machine translation on low-resource African languages, which primarily assesses cross-lingual transfer learning and generalization of machine translation metrics across a wide range of under-resourced languages. In this paper, we analyze the submissions to WMT 2024 Metrics Shared Task. Our findings indicate that language-specific adaptation, cross-lingual transfer learning, and larger language model sizes contribute significantly to improved metric performance. Moreover, supervised models with relatively moderate sizes demonstrate robust performance, when augmented with specific language adaptation for low-resource African languages. Finally, submissions show promising results for language pairs including Darija-French, English-Egyptian Arabic, and English-Swahili. However, significant challenges persist for extremely low-resource languages such as English-Luo and English-Twi, highlighting areas for future research and improvement in machine translation metrics for African languages.

1 Introduction

Recent machine translation (MT) research has scaled dramatically, encompassing hundreds of languages, including many under-resourced ones (Fan et al., 2021a; NLLB-Team et al., 2022; Bapna et al., 2022; Kudugunta et al., 2023). However, accurately measuring MT quality in low-resource languages remains challenging. Conventional metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and chrF (Popović, 2015), which rely on n-gram matching, often fail to capture deeper semantic similarities (Zhang et al., 2020; Rei et al., 2020; Sai B et al., 2023).

Newer approaches include embedding-based metrics like BERTScore (Zhang et al., 2020) and learned metrics such as COMET (Rei et al., 2020), which have shown promise in more accurately evaluating translations across diverse languages. However, the application of these neural-based metrics to under-resourced languages continues to face significant challenges (Wang et al., 2024), highlighting ongoing areas of research in multilingual MT evaluation. These challenges include: (1) data scarcity impeding metric development, (2) complexity in annotation guidelines challenging non-expert evaluators, and (3) limited language model coverage restricting applicability, which underscore the need for continued innovation in MT evaluation methods, particularly for under-resourced African languages.

In response to these challenges, Wang et al. (2024) have introduced AFRIMTE, a human evaluation dataset focusing on MT adequacy and fluency for 13 typologically diverse African languages. This dataset addresses the data scarcity issue and employs simplified MQM evaluation guidelines tailored for non-expert translators, thus tackling two of the primary challenges in this field. Moreover, the authors establish benchmark systems for MT Evaluation and reference-free Quality Estimation (QE) by leveraging transfer learning techniques. These techniques draw from existing, well-resourced Direct Assessment (Graham et al., 2013) (DA) data and utilize an African-centric multilingual pre-trained language model, thereby addressing the challenge of limited language model coverage for African languages.

Building on this work, the WMT 2024 Metrics Shared task incorporates the translation adequacy test set from AFRIMTE as a challenge set. This inclusion aims to evaluate the capabilities of metric systems for machine translation on low-resource African languages, primarily assessing the cross-lingual transfer learning ability and generalization of these systems across a wide range of under-resourced African languages.

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 505–516
November 15–16, 2024.

Our examination of task submissions has yielded several key findings in the development of machine translation metrics for African languages. We observed that language-specific adaptation, cross-lingual transfer learning, and increased language model sizes contribute to significant improvements in metric performance. Even supervised models of relatively modest scale can achieve robust results when augmented with language adaptation techniques. In addition, our analysis reveals promising outcomes for certain language pairs, such as Darija-French, English-Egyptian Arabic, and English-Swahili. However, persistent challenges remain evident in extremely low-resource languages like English-Luo and English-Twi. These disparities underscore critical areas requiring further investigation and highlight the need for targeted research in developing effective metrics across the diverse linguistic landscape of Africa.

2 AfriMTE

AfriMTE (Wang et al., 2024) focuses on the dev and devtest subsets of the FLORES-200 dataset (NLLB-Team et al., 2022). It covers 13 language pairs (LPs), primarily focusing on African languages with English, plus Darija-French and a control pair of English-French. In details, there are Darija-French (ary-fr), English-Egyptian Arabic (en-arz), English-French (en-fr), English-Hausa (en-hau), English-Igbo (en-ibo), English-Kikuyu (en-kik), English-Luo (en-luo), English-Somali (en-som), English-Swahili (en-swh), English-Twi (en-twi), English-isiXhosa (en-xho), English-Yoruba (en-yor), and Yoruba-English (yor-en). The annotations were also extended on domain-specific translations for English-Yoruba.

The MT outputs were generated using two open-source MT engines: NLLB-200 (600M) (NLLB-Team et al., 2022) and M2M-100 (418M) (Fan et al., 2021b). Most language pairs use NLLB-200, except for English-French and English-Swahili, which use M2M-100 due to their exceptionally high translation quality based on NLLB-200. The authors noted that while some language pairs like English-isiXhosa showed high overall quality, minor errors at the word level were still present.

AfriMTE initially provides both fine-grained word-level error annotations and sentence-level Direct Assessment scoring for translation adequacy and fluency. For the WMT 2024 Metrics Shared Task, we utilize the adequacy test set from AfriMTE as the African Challenge set to evaluate the sentence-level scoring performance of submitted metrics, focusing specifically on the FLORES-200 subsets within the dataset. Table 1 presents the counts of translation annotations in this challenge set. Due to the limited sizes of annotations for individual language pairs, we merge test data from all LPs into a single African-centric dataset to enhance evaluation significance for MT evaluation and reference-free quality estimation (QE) metrics. However, recognizing that different LPs may have varying score ranges, potentially favoring metrics that correlate with these distributions more than actual quality, we also report metric performance on each LP separately. This approach balances the need for statistical robustness with LP-specific insights.

LP	Test #	LP	Test #
ary-fr	187	en-som	226
en-arz	250	en-swh	157
en-fr	250	en-twi	247
en-hau	240	en-xho	243
en-ibo	120	en-yor	239
en-kik	202	yor-en	212
en-luo	242		
Total: 2815 annotations			

Table 1: Counts of adequacy annotations for each language pair (LP) in the test set of AfriMTE.

3 Metrics

The WMT 2024 Metrics Shared Task received various metric submissions from both task organizers and participants. Our analysis will concentrate on the baseline metrics provided by the task organizers and the primary and contrastive metrics submitted by the participants.

3.1 Baselines

The baseline metrics for MT evaluation include BLEU (Papineni et al., 2002), chrF (Popović, 2015), spBLEU (Fan et al., 2021a), prism-Ref (Thompson and Post, 2020), YiSi-1 (Lo, 2019), COMET-22 (Rei et al., 2022a), BLUERT-20 (Sellam et al., 2020), and BertScore (Zhang et al., 2019). For reference-free quality estimation, the baseline metric is CometKiwi (Rei et al., 2022b). Additionally, we include AfriCOMET and AfriCOMET-QE for comparison, which are the African extensions of COMET-22 (Rei et al., 2022a) and CometKiwi (Rei et al., 2022b) proposed by Wang et al. (2024). They employ transfer learning from well-resourced DA data and utilize an African-centric multilingual pre-trained encoder, AfroXLMR (Alabi et al., 2022), to build MT evaluation and QE models for African languages.

3.2 Submissions from Participants

The metrics submitted by participants for MT evaluation include XCOMET (Guerreiro et al., 2023), METRICX-24 and METRICX-24-HYBRID (Juraska et al., 2024)[1], chrF-S (Mukherjee and Shrivastava, 2024), METAMETRICS-MT (Anugraha et al., 2024), damonmonli, and monmonli[2]. For reference-free QE, the submitted metrics are XCOMET-QE (Guerreiro et al., 2023), METRICX-24-QE and METRICX-24-HYBRID-QE (Juraska et al., 2024)[3], QE model of METAMETRICS-MT (Anugraha et al., 2024), GEMBA-ESA (Kocmi and Federmann, 2023), and XLsimMQM (Mukherjee and Shrivastava, 2023). Details of all metrics can be found in Freitag et al. (2024).

3.3 AfriCOMET-1.1 and AfriCOMET-QE-1.1

In the ongoing efforts to enhance performance on African languages, we explore the use of a more advanced African pre-trained encoder. Specifically, we re-train AfriCOMET and AfriCOMET-QE using AfroXLMR-76 (Adelani et al., 2024) and conduct the training in single-task learning mode (Wang et al., 2024).

AfroXLMR-76 (Adelani et al., 2024) is an enhanced version of AfroXLMR (Alabi et al., 2022), which itself was a multilingual adaptation of the XLM-R-large model for 20 widely spoken African languages (each with at least 50MB of data). AfroXLMR-76 scales the language coverage up to 76 languages, including 61 languages with at least 10MB of data and an additional 15 languages with less than 10MB. To address the scarcity of monolingual data for some African languages, Adelani et al. (2024) proposed to generate synthetic parallel sentences by translating an English news commentary dataset (Kocmi et al., 2022) using the NLLB (600M) model.This expanded language coverage and increased training data volume have resulted in AfroXLMR-76 outperforming its predecessor, AfroXLMR, on the SIB-200 topic classification dataset for African languages (Adelani et al., 2024).

Metrics	Pearson	Spearman	Kendall
METRICX-24*	0.5188	0.3949	0.2714
AfriCOMET-1.1*	0.5117	0.4129	0.2865
AfriCOMET-1.0	0.4821	0.3857	0.2675
METRICX-24-HYBRID	0.4764	0.3844	0.2640
METAMETRICS-MT	0.3934	0.3429	0.2360
COMET-22	0.3674	0.2835	0.1943
YiSi-1	0.3058	0.2453	0.1666
chrF-S	0.3121	0.2332	0.1584
chrF	0.2833	0.2193	0.1492
BERTScore	0.2959	0.1834	0.1248
BLEURT-20	0.2284	0.2225	0.1492
XCOMET	0.2224	0.2119	0.1451
spBLEU	0.2159	0.2052	0.1388
monmonli	0.2022	0.1713	0.1152
damonmonli	0.2007	0.1690	0.1138
BLEU	0.1863	0.1897	0.1282
PrismRefMedium	0.1149	0.1799	0.1202
PrismRefSmall	0.1058	0.1642	0.1099

Table 2: Segment-level correlation coefficients of **MT evaluation** metrics on the entire AFRIMTE. Metrics marked with * are ranked first based on the Perm-Input hypothesis test (Deutsch et al., 2021).

We refer to the original models using AfroXLMR as AfriCOMET-1.0[4] and AfriCOMET-QE-1.0[5], while the new versions leveraging AfroXLMR-76 are called AfriCOMET-1.1[6] and AfriCOMET-QE-1.1[7], respectively.

4 Analysis

This section presents a comprehensive analysis of the metrics outlined in Section 3. Our evaluation framework is structured around two primary components. First, we assess segment-level performance by examining the correlation between metric scores and human DA scores. This assessment involves analyzing correlation coefficients on the entire mixed African Challenge set and calculating weighted average correlation coefficients across various language pairs. Second, we conduct a language-specific analysis by computing average correlation coefficients for each individual language pair across all metric systems.

[1]METRICX-24 is the contrastive system to METRICX-24-HYBRID

[2]The monmonli is the contrastive system to damonmonli.

[3]METRICX-24-QE is the contrastive system to METRICX-24-HYBRID-QE

[4]https://huggingface.co/masakhane/africomet-stl

[5]https://huggingface.co/masakhane/africomet-qe-stl

[6]https://huggingface.co/masakhane/africomet-stl-1.1

[7]https://huggingface.co/masakhane/africomet-qe-stl-1.1

Metrics	Pearson	Spearman	Kendall
METRICX-24*	0.6269	0.4833	0.3455
METRICX-24-HYBRID	0.5972	0.4695	0.3351
METAMETRICS-MT	0.5295	0.4726	0.3368
AfriCOMET-1.1	0.5399	0.4363	0.3097
AfriCOMET-1.0	0.5260	0.4261	0.3027
XCOMET	0.4108	0.4045	0.2874
COMET-22	0.4513	0.3432	0.2430
YiSi-1	0.4233	0.3125	0.2182
BLEURT-20	0.3604	0.3428	0.2396
BERTScore	0.3997	0.2933	0.2049
chrF-S	0.3763	0.3025	0.2106
damonmonli	0.3627	0.3013	0.2100
chrF	0.3593	0.2955	0.2053
monmonli	0.3215	0.2877	0.1991
PrismRefMedium	0.2389	0.2978	0.2053
PrismRefSmall	0.2250	0.2868	0.1984
spBLEU	0.2585	0.2515	0.1733
BLEU	0.2394	0.2457	0.1691

Table 3: Segment-level weighted average correlation coefficients of **MT evaluation** metrics, averaged across language pairs on AFRIMTE, with weights based on the size of each language pair group. The metric marked with * ranks first based on the average of Pearson, Spearman, and Kendall correlation coefficients.

4.1 Segment-level Averaged Correlation

For both MT evaluation and reference-free QE tasks, we assess the metric performance using three widely adopted correlation coefficients: Pearson, Spearman-rank, and Kendall-rank. These coefficients measure the correlation between metric scores and human DA scores, each capturing different aspects of the relationship (Deutsch et al., 2023). To validate the statistical significance of our results, we additionaly employ the Perm-Input hypothesis test (Deutsch et al., 2021), which is conducted with 200 re-sampling runs and a significance level of $p = 0.05$, producing rankings of the various automatic metrics based on their performance.

4.1.1 MT Evaluation Metric

We present the segment-level correlation coefficients of MT evaluation metrics on the entire AFRIMTE test set in Table 2 and the weighted average correlation coefficients across various language pairs in Table 3. Detailed Pearson, Spearman-rank, and Kendall-rank correlations of baseline metrics and primary submissions for each language pair are shown in Figures 3, 4, and 5 of Appendix A.

For MT evaluation, Table 2 provides valuable insights into evaluation metrics' performance on the African Challenge Set. Generally, Pearson correlations are generally higher than Spearman and Kendall, with rankings remaining largely consistent across correlation types. The top-performing metrics—METRICX-24, AfriCOMET-1.1, AfriCOMET-1.0, and METRICX-24-HYBRID—are all based on pre-trained multilingual large language models (LLMs) and utilize supervised learning. These metrics consistently outperform other types across all correlation coefficients. METRICX-24 and AfriCOMET-1.1 emerge as the best performers, statistically indistinguishable from each other. The improved performance of AfriCOMET-1.1 over its predecessor suggests ongoing enhancements in these LLM-based metrics. It is evident that African-centric LLM-based metrics (AfriCOMET variants) perform exceptionally well, highlighting the importance of language-specific fine-tuning for low-resource African languages.

Moreover, the weighted average correlation results presented in Table 3 offer additional valuable insights. METRICX-24 still emerges as the top-performing metric, achieving the highest correlation with human judgments across all three correlation coefficients (Pearson: 0.6269, Spearman: 0.4833, Kendall: 0.3455). Its hybrid variant, METRICX-24-HYBRID, follows closely, suggesting the robustness of this metric family. METAMETRICS-MT shows strong performance, ranking third overall with high correlation coefficients. As an ensemble method, it selectively combines complementary metrics, proves effective for African languages despite these metrics being trained on general WMT data. In addition, AfriCOMET-1.1 and its predecessor AfriCOMET-1.0 show robust performance indicating their effectiveness for African language pairs.

Traditional metrics like BLEU and its variant spBLEU demonstrate relatively weak correlations, reinforcing the need for more advanced metrics in evaluating MT quality for African languages. Interestingly, some widely-used metrics such as BERTScore and BLEURT-20 show moderate performance, outperforming traditional metrics but falling behind the top-performing ones. The consistent ranking across different correlation coefficients suggests a reliable performance hierarchy among these metrics. However, the overall moderate correlation values (mostly below 0.5 for Spearman and Kendall) highlight the ongoing challenges in accurately evaluating MT quality for African languages.

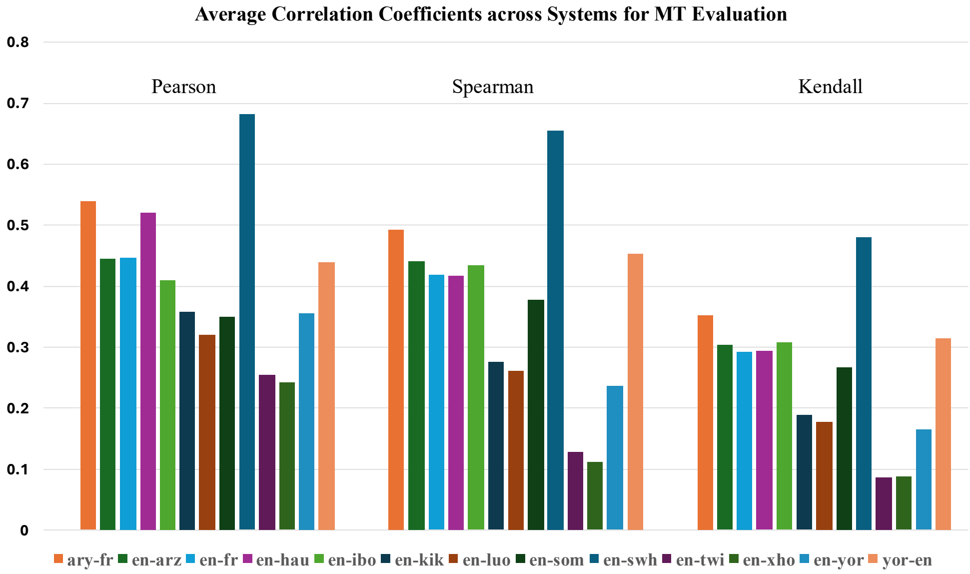

Figure 1: Average correlations across MT evaluation metrics for each language pair.

Metrics	Pearson	Spearman	Kendall
METRICX-24-QE*	0.4857	0.3810	0.2616
AfriCOMET-QE-1.1*	0.4760	0.3961	0.2747
METRICX-24-HYBRID-QE	0.4337	0.3594	0.2464
GEMBA-ESA	0.4033	0.3300	0.2427
METAMETRICS-MT	0.3781	0.3004	0.2050
AfriCOMET-QE-1.0	0.3496	0.2524	0.1729
CometKiwi-XXL	0.2149	0.1814	0.1254
XCOMET-QE	0.1717	0.1528	0.1042
CometKiwi	0.1685	0.1259	0.0838
XLsimMQM	0.0886	0.0925	0.0619

Table 4: Segment-level correlation coefficients of **QE** metrics on AFRIMTE. Metrics marked with * are ranked first based on the Perm-Input hypothesis test (Deutsch et al., 2021).

Metrics	Pearson	Spearman	Kendall
METRICX-24-QE*	0.5790	0.4383	0.3117
METRICX-24-HYBRID-QE	0.5530	0.4289	0.3048
AfriCOMET-QE-1.1	0.4905	0.4117	0.2900
GEMBA-ESA	0.4624	0.3793	0.2900
METAMETRICS-MT	0.5010	0.3610	0.2528
AfriCOMET-QE-1.0	0.4774	0.3743	0.2628
CometKiwi-XXL	0.3709	0.3428	0.2417
XCOMET-QE	0.3087	0.3290	0.2317
CometKiwi	0.3301	0.2914	0.2046
XLsimMQM	0.1548	0.1817	0.1256

Table 5: Segment-level weighted average correlation coefficients of **QE** metrics, averaged across language pairs on AFRIMTE, with weights based on the size of each language pair group. The metric marked with * ranks first based on the average of Pearson, Spearman, and Kendall correlation coefficients.

4.1.2 Quality Estimation as a Metric

QE presents a more challenging and purely cross-lingual task, making its investigation essential. Tables 4 and 5 presents the segment-level correlation coefficients of QE metrics on the entire AFRIMTE and weighted average correlations across language pairs. Detailed Pearson, Spearman-rank, and Kendall-rank correlations of baseline metrics and primary submissions for each language pair are shown in Figures 6, 7, and 8 of Appendix A.

Comparing results in Tables 2 and 4, and results in Tables 3 and 5, we have observed significant performance gaps between MT evaluation models and their QE counterparts. This is evident when comparing specific versions, such as the differences between METRICX-24 and METRICX-24-QE, XCOMET and XCOMET-QE, as well as AfriCOMET-1.1 and AfriCOMET-QE-1.1. These disparities underscore the increased complexity of the QE task, which requires assessing translation quality without access to reference translations.

Tables 4 and 5 reveal the superior performance of LLM-based supervised-learning metrics in the QE task. Specifically, METRICX-24-QE and AfriCOMET-QE-1.1 emerge as the top-performing metrics on the entire AFRIMTE test set (Table 4). These metrics demonstrate statistically indistinguishable performance, as confirmed by the Perm-Input hypothesis test. Furthermore, in the weighted average correlation across different language pairs (Table 5), METRICX-24-QE consistently outperforms other approaches. This trend in QE metrics mirrors the pattern observed in MT evaluation metrics, underscoring the effectiveness of LLM-based supervised-learning approaches in both contexts for African languages. Additionally, METAMETRICS-MT, as a meta-metric, continues to show strong performance, further validating the effectiveness of ensemble methods in addressing the complexities of African language evaluation. Another LLM-based metric, GEMBA-ESA, which employs a two-step approach: first collecting MQM error spans, and then assigning the final score also demonstrates robust performance, further highlighting the potential of LLM-based techniques in QE tasks for African languages. However, supervised QE metrics such as CometKiwi, CometKiwi-XXL, and XCOMET-QE show relatively poor performance, suggesting they might not be well-suited for African languages without specific language adaptation.

4.1.3 Language Adaptation, Cross-lingual Transfer, and Model Size as Key Factors in Metric Performance

Our analysis on the baseline and task submissions reveals that language-specific tuning, cross-lingual transfer learning, and model size are crucial factors in MT evaluation and Quality Estimation.

The top-performing systems demonstrate these principles in various ways. METRICX-24 systems, based on mT5-XXL (Xue et al., 2020), cover a wide range of languages, including several African languages such as Hausa, Igbo, Somali, Swahili, Xhosa, Yoruba, and Zulu. In contrast, AfriCOMET models use African-enhanced masked language models (AfroXLMR and AfroXLMR-76) with well-resourced DA training data, showcasing the benefits of language-specific adaptation. Both METRICX-24 and AfriCOMET variants employ supervised training and cross-lingual transfer learning, proving effective for low-resource language scenarios. The impact of model size is evident, with AfriCOMET variants (560 million parameters) and METRICX-24 (13 billion parameters) both achieving strong results. While METRICX-24's larger size contributes to its superior performance, AfriCOMET's performance demonstrates that well-adapted smaller models can also yield robust results.

Moreover, the excellent performance of METAMETRICS-MT underscores the potential of ensembling robust metrics to create effective meta-metrics. The promising results of GEMBA-ESA further highlight the effectiveness of LLM-based prompting techniques in this domain. These findings collectively emphasize the potentials of model ensemble and innovative LLM prompting strategies in developing effective MT evaluation and QE metrics, particularly for low-resource languages.

4.2 Language-Specific Performance: Average Correlations across Metrics

To investigate model performance on specific language pairs, we calculate the average correlation coefficients for each individual language pair across all metric systems, providing insights into how well metrics perform for specific language pairs. Results are shown in Figure 1 and 2.

4.2.1 Performance on MT Evaluation

Figure 1 depicting the average correlation coefficients across metric submissions for MT evaluation reveals significant variations in metric performance on different language pairs. Consistently across all pairs, Pearson correlation shows the highest values, followed by Spearman and then Kendall, suggesting stronger linear relationships between human and metric scores compared to monotonic or ordinal relationships. English-Swahili (en-swh) and Darija-French (ary-fr) demonstrate the highest correlations across all three metrics, likely due to their status as more resource-rich or commonly studied pairs. In contrast, English-Luo (en-luo), English-Twi (en-twi), and English-isiXhosa (eng-xho) exhibit the lowest correlations, indicating particular challenges for MT evaluation in these language pairs.

4.2.2 Performance on QE

A consistent pattern emerges in the QE task (Figure 2) where Pearson correlations generally show the highest values. Language pair performance is notably similar across both figures, with resource-rich pairs like English-Swahili (en-swh) consistently demonstrating higher correlations, while extremely low-resource pairs such as English-Luo (en-luo) and English-Twi (en-twi) show persistently lower correlations. Interestingly, some language pairs show improved relative performance in QE compared to MT Evaluation. For example, English-Egyptian Arabic (en-arz) and English-Hausa (en-hau) demonstrate better results in QE, possibly indicating their suitability for reference-free evaluation methods.

4.2.3 Some Special Cases

Contrary to expectations, English-French (en-fr) does not emerge as the top-performing language

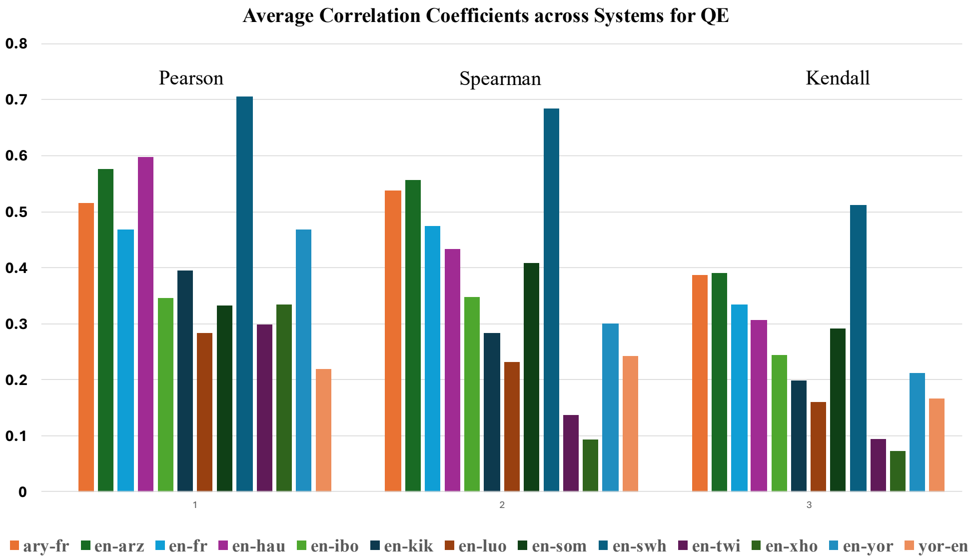

Figure 2: Average correlations across QE metrics for each language pair.

pair in either the MT evaluation or the QE task. This surprising result might be attributed to two factors. First, as illustrated in Table 7 of Wang et al. (2024), there is a scarcity of supervised DA training datasets for English-French. Second, the performance may be affected by the "curse of multilinguality" (Pfeiffer et al., 2022), a phenomenon where model performance on high-resource languages can degrade when the pre-trained model is fine-tuned and enhanced with data from multiple low-resource languages, in this case, African languages.

Another noteworthy case is English-isiXhosa (en-xho). As previously observed, English-isiXhosa translations demonstrated high overall sentence-level quality (median DA: 100 according to Wang et al. (2024)) , with only minor errors at the word level. This characteristic makes it particularly challenging to differentiate and rank translation quality. Consequently, the relatively lower performance of Spearman and Kendall for English-isiXhosa is expected.

5 Conclusion

In conclusion, our analysis on submissions to the AfriMTE challenge set of WMT 2024 Metrics Shared Task for African languages reveals that LLM-based supervised-learning metrics, especially those with African-centric tuning, consistently outperform traditional and other neural-based approaches in both MT evaluation and Quality Estimation tasks. Language-specific adaptation, cross-lingual transfer learning, and larger model sizes contribute significantly to improved metric performance. However, challenges persist for extremely low-resource languages such as Luo and Twi. Our analysis also highlights unexpected performance patterns in certain language pairs, including English-French and English-isiXhosa, demonstrating the complexities of evaluating machine translation across diverse African languages.

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun 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.

David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya, and Genta Indra Winata. 2024. Metametrics-MT: Tuning machine translation metametrics via human preference calibration. In *Proceedings of the Ninth Conference on Machine Translation*, 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.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi N. Baljekar, Xavier García, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Z. Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages. *ArXiv*, abs/2205.03983.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929.

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. 2021a. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021b. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

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*, USA. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*.

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*, USA. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. *arXiv preprint arXiv:2310.13988*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier García, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *ArXiv*, abs/2309.04662.

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.

Ananya Mukherjee and Manish Shrivastava. 2023. MEE4 and XLsim: IIIT HYD's Submissions for WMT23 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2024. chrf-s: Semantics is all you need. In *Proceedings of the Ninth Conference on Machine Translation*, USA. 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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L.

Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

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.

Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. *arXiv preprint arXiv:2205.06266*.

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.

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.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Jiayi Wang, David 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, Shamsuddeen Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Mohamed, Hassan Ayinde, Oluwabusayo Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Toadoum Sari Sakayo, Lyse Naomi Wamba, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Iro, Saheed Abdullahi, Stephen Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Ogbu, Sam Ochieng', Verrah Otiende, Chinedu Mbonu, Yao Lu, and Pontus Stenetorp. 2024. 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.

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. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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

A Appendix

Detailed Pearson, Spearman-rank, and Kendall correlation coefficients of MT evaluation and QE metrics for each language pair are shown in Figures 3, 4, 5, 6, 7, and 8 accordingly.

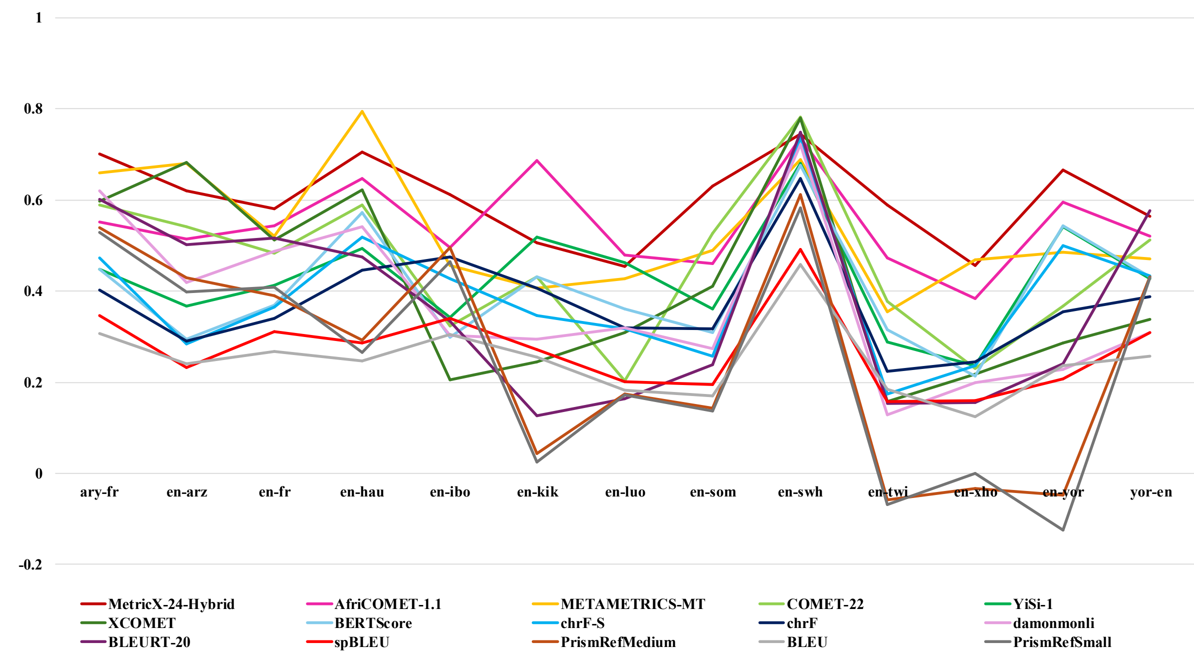

Figure 3: Pearson Correlations of MT Evaluation Metrics for each language pair.

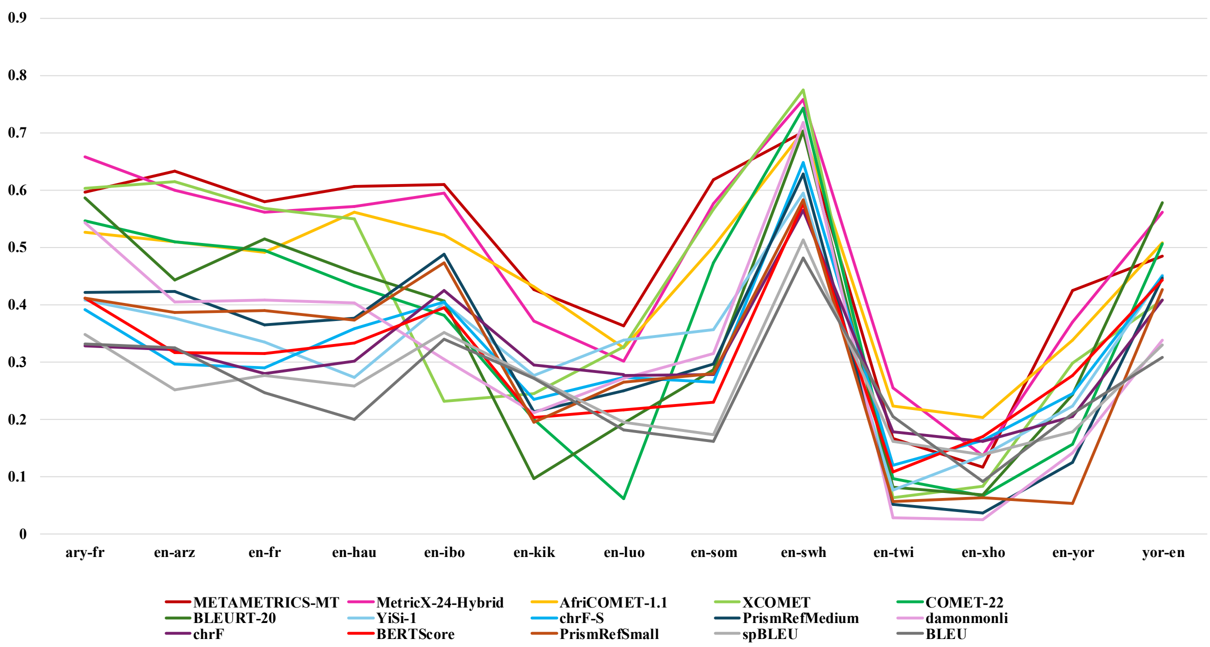

Figure 4: Spearman-rank Correlations of MT Evaluation Metrics for each language pair.

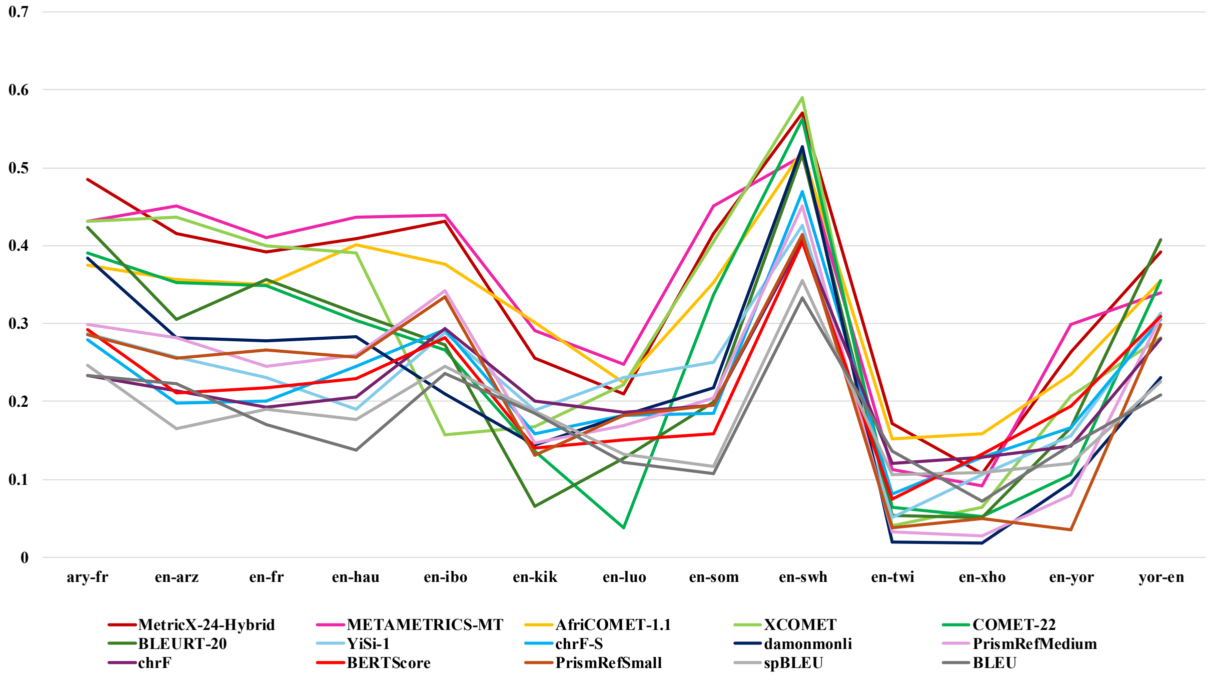

Figure 5: Kendall-rank Correlations of MT Evaluation Metrics for each language pair.

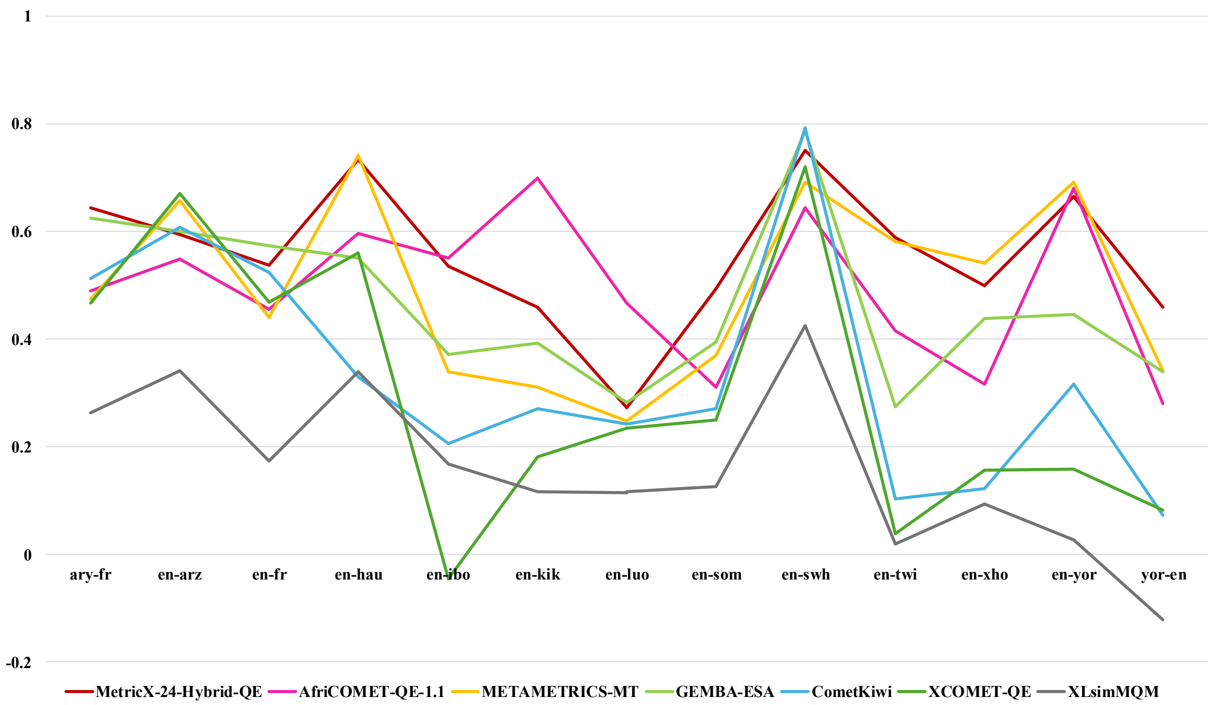

Figure 6: Pearson Correlations of QE Metrics for each language pair.

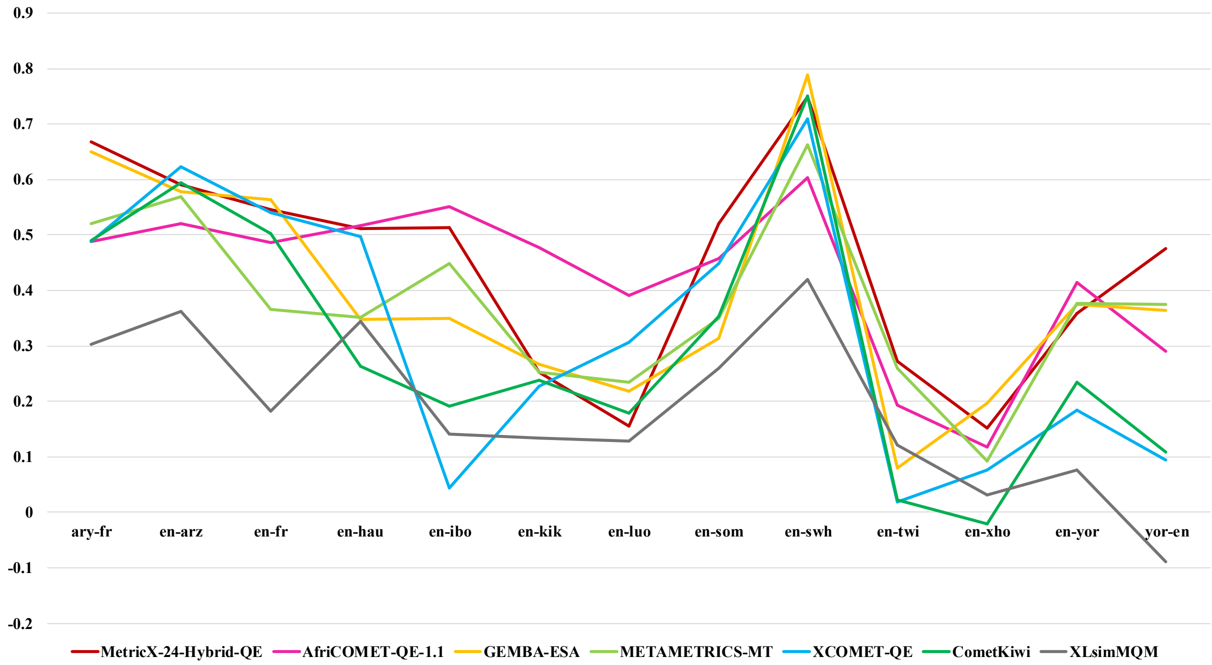

Figure 7: Spearman-rank Correlations of QE Metrics for each language pair.

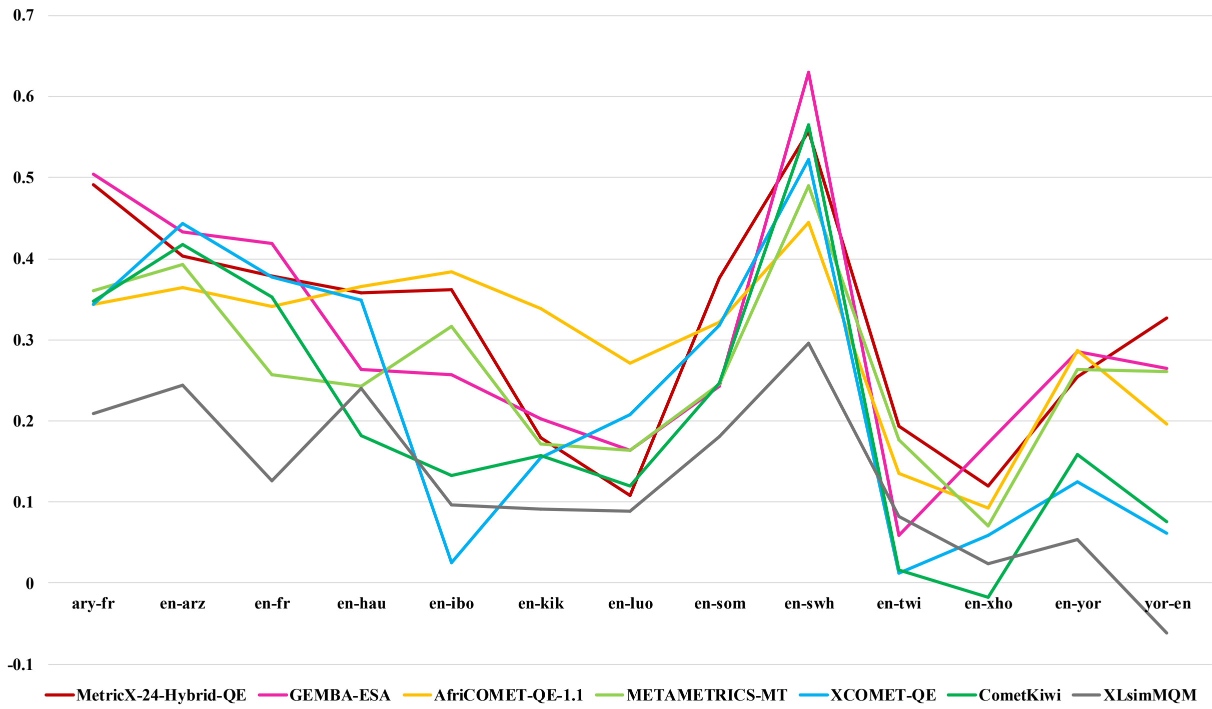

Figure 8: Kendall-rank Correlations of QE Metrics for each language pair.