Metric Score Landscape Challenge (MSLC23): Understanding Metrics' Performance on a Wider Landscape of Translation Quality

Chi-kiu Lo 羅致翹 **Samuel Larkin** **Rebecca Knowles**
Digital Technologies Research Centre
National Research Council Canada (NRC-CNRC)
{chikiu.lo,samuel.larkin,rebecca.knowles}@nrc-cnrc.gc.ca

Abstract

The Metric Score Landscape Challenge (MSLC23) dataset aims to gain insight into metric scores on a broader/wider landscape of machine translation (MT) quality. It provides a collection of low- to medium-quality MT output on the WMT23 general task test set. Together with the high quality systems submitted to the general task, this will enable better interpretation of metric scores across a range of different levels of translation quality. With this wider range of MT quality, we also visualize and analyze metric characteristics beyond just correlation.

1 Introduction

Under time and human resource constraints, automatic metrics are often used as a proxy of manual evaluation for machine translation (MT) quality. The WMT Metrics shared task evaluates how well a variety of automatic metrics correspond to human judgments of MT quality, as evaluated on the WMT General (formerly News) shared task data. Those MT systems being evaluated are typically high-performing systems, especially for high-resource language pairs. However, in practice, the lessons learned are applied to a broader range of systems in development, including low-resource and low-quality output.

This challenge set[1] aims to gain insight into metric scores across a broader MT quality landscape. It provides a collection of low- to medium-quality MT output on the WMT23 general task test set. This serves several purposes. Together with the high quality systems submitted to the general task, this will enable more thorough understanding of metric scores across a range of different levels of translation quality: useful knowledge for researchers considering applying these metrics to lower-resource language pairs or lower-performing domains. This challenge set also allows us to explore metric characteristics beyond just correlation, which has been a main focus of past work. By expanding the range of MT quality analyzed, we shed light on some unexpected or under-explored properties of metrics, such as metrics that can distinguish between high quality systems but are not able to differentiate different levels of MT quality on the lower end of the quality scale (or vice versa) and metrics that use their space of scores in very different ways (e.g., discretized, or with specific score ranges with particularly large numbers of ties).

We focus on four language pairs: Chinese→English (ZH→EN), Hebrew↔English (HE↔EN), and English→German (EN→DE). Three of these correspond to the focus languages of the WMT 2023 Metrics shared task (EN→DE, HE→EN, ZH→EN), and they also cover several language families and aspects of translation evaluation (i.e., the paragraph-level evaluation of EN→DE), as well as including a sentence-level out-of-English direction (EN→HE). We combine source and reference data from the news portion of the WMT 2023 General MT task test sets with our challenge set, the low- and medium-quality MT output that we generated to cover a range of MT quality.

We begin by describing the training data (Section 3.1) and models (Section 3.2) used in for constructing our challenge set (Section 3.3). We also briefly describe the additional data (Section 4) and metrics (Section 5) analyzed. In Section 6 we analyze the distribution of different metrics over the challenge set. We find that some metrics exhibit strikingly different characteristics on the low-quality systems as compared to the systems submitted to WMT, while others exhibit unexpected characteristics (e.g., large numbers of tied scores) that would not have been apparent from standard correlational analysis or from high-quality WMT submitted systems alone. We conclude by arguing

[1] Available at https://github.com/nrc-cnrc/MSLC23

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 776–799
December 6–7, 2023.

that examining metric characteristics and performance over a wider landscape of MT quality—or indicating clearly when a metric has only been tested on high-quality MT—is an important factor for researchers to consider when building, presenting, and applying new metrics (especially if those metrics will be applied to lower-quality outputs).

2 Related Work

Przybocki et al. (2009) outlined four objectives in the search for new and improved automatic MT evaluation metrics: 1) "high correlation with human assessments of translation quality"; 2) "applicable to multiple target languages"; 3) "ability to differentiate between systems of varying quality" and finally, 4) "intuitive interpretation". Over the years, the WMT Metrics shared tasks (Callison-Burch et al., 2007; Bojar et al., 2017b; Freitag et al., 2021, 2022, i.a.) focused mainly on evaluating MT evaluation metrics on the first two objectives.

Many other research efforts on meta-evaluation of metrics also focused on their ability to correlate with human judgment. Graham and Baldwin (2014) introduced Williams' significance tests for understanding the confidence of the correlation analysis. Mathur et al. (2020) pointed out that Pearson's correlation is sensitive to outliers and proposed to remove outliers in Pearson's correlation analysis at system level. Kocmi et al. (2021) proposed to use pairwise accuracy to evaluate metrics based on whether the metric's pairwise rankings of two systems agrees with human pairwise rankings. Deutsch et al. (2023) introduced a tie calibration procedure enabling fair comparison between metrics that do and do not predict ties for pairwise accuracy analysis at the segment level. Marie (2022) and Lo et al. (2023) studied the relationship of metrics' score differences and statistical significance of ranking decision. Notably, these works are mostly based on the data released by WMT Metrics shared task. That means the translation output scored by the metrics in these work were generated by the participants of the WMT News/General Translation shared task, typically consisting of high-quality MT output.

There is growing interest in understanding metric performance beyond correlation. Moghe et al. (2023) note that neural metrics are not interpretable at the segment level across different language pairs. The WMT Metrics shared task introduced the challenge sets subtask (Freitag et al., 2021, 2022) to challenge metrics on particular translation errors, including negation and polarity, word/sentence addition/omission, tokenization, punctuation, numeric expression, casing number swapping, spelling, etc., in order to shed light on metric strengths and weaknesses. The challenge sets created by Macketanz et al. (2018); Avramidis et al. (2020); Avramidis and Macketanz (2022) were more linguistically motivated and covers more than 100 phenomena, including tenses, relative clauses, idioms, focus particles, etc. The ACES challenge set (Amrhein et al., 2022) covers 146 translation directions and 68 types of errors, ranging from simple perturbations to more complex errors based on discourse and real-world knowledge. The SMAUG challenge set (Alves et al., 2022) and the HWTSC challenge set (Chen et al., 2022) focused on the robustness of metrics on translation errors involving named entities, numeric/date/time entities, etc.

We note that even as MT evaluation metrics become better at correlating with human judgment on translation quality for high-quality MT systems, metric performance may be untested on low- to medium-quality MT output. Hence, we design the MSLC23 challenge set to gain insights of metric behavior on a more complete landscape.

3 Challenge Set

The challenge set consists of data translated by MT systems of varying quality. We describe the training data used to build these systems as well as the MT models.

3.1 Training Data

To build the lower-quality MT systems that we analyze in this work, we use standard WMT datasets from WMT 2023 (Kocmi et al., 2023) for EN→DE and HE↔EN and from WMT 2017 (Bojar et al., 2017a) for ZH→EN. For EN→DE and ZH→EN, we used the *newstest2020* data as our validation set. For HE↔EN, we used a random sample of 2000 lines, ensuring no overlap between sentence pairs in the training and validation set. For full details of training data, see Appendix A. Appendix B describes the preprocessing and subword segmentation performed.

3.2 MT Models

We build two main types of systems: baselines and pseudo-low-resource systems. All systems were

built using Sockeye-3.1.31 (Hieber et al., 2022), commit 13c63be5, with PyTorch-1.12.1 (Paszke et al., 2019). For more details on parameters and training, see Appendix C.

The baselines are standard Transformer models trained over the available data, but without any additional components (e.g., backtranslation, factors, tagging, etc.). The pseudo-low-resource systems are produced using subsets of the training data, to simulate lower-resource settings (see Appendix D for details). We checkpoint all systems frequently so that we can use output at various levels of training as representative of different levels of quality.

We note that the EN→DE 2023 shared task is performed at the paragraph level. In our work we do not perform paragraph-level MT; instead we use as a baseline sentence segmentation, translation of the individual sentences, and concatenation back into paragraphs of the resulting translation output.

3.3 Translation Output

We use the news data portions of each of the 2023 General MT task test sets for these language pairs. This consists of 139 paragraphs for EN→DE (translated by 12 different systems), 516 lines for EN→HE (translated by 6 different systems), 1558 lines for HE→EN (translated by 6 different systems), and 763 lines for ZH→EN (translated by 6 different systems).

We use checkpoints from each of the systems we built to produce the low- and medium-quality MT output. For ZH→EN and HE↔EN, all checkpoints were selected from the baseline systems; for EN→DE, they were selected from the baseline system as well as the 50k, 200k, and 400k pseudo-low-resource systems. These checkpoints were selected to cover a range of BLEU scores from less than 1 to between 20 and 30 (shown in Appendix E, Tables 11 and 12; we assign the selected systems the letters A through F, or through L in the case of EN→DE, with A being the lowest quality system in all cases),[2] and were then spot-checked manually to confirm that they did generally appear to represent incremental (but noticeable) improvements in quality. We note that we did not perform a full or extensive manual evaluation, and as such cannot make claims of statistically significant human judgment differences between the checkpoints. Another potential limitation of this choice to select checkpoints is that they may be more similar to one another than separately-trained systems would be (cf. the benefits of ensembling diverse sets of systems or the potential minor drawbacks of ensembling checkpoints rather than separately trained models in Farajian et al. (2016); Sennrich et al. (2016), i.a.). Nevertheless, we expect this should provide some coverage of low- to mid-quality MT for scoring by various metrics.

[2]Computed with sacreBLEU (Post, 2018) with signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1 For HE↔EN, the systems were selected based on BLEU scores computed with refA as the reference, as refB had not yet been released. The range of BLEU scores remains similar, and we present all other HE↔EN results based on scores with refB.

4 General MT Submissions

Our challenge set has aimed to cover the range of low-quality MT systems, but to obtain a fuller picture, we also include the metric scores assigned to the systems submitted to the WMT2023 General MT Task (Kocmi et al., 2023) in our analysis. For these systems, we have human annotation scores in the form of multidimensional quality metrics (MQM; Burchardt, 2013) scores for EN→DE, HE→EN, and ZH→EN.

5 Metrics

There are dozens of metrics submitted by the task organizers and participants in WMT23 Metrics shared task. Under the time and space limitations, we only examine the baseline metrics submitted by the task organizers and the primary metrics submitted by the participants. Due to the random shuffling of items in the challenge sets before their delivery to the scorers, we can only examine metrics that produce scores at the segment level, as the system-level scores returned do not correspond to the underlying systems in our datasets. We describe the metrics included in this work in Appendix F.

6 Analysis

We are interested in metric performance and characteristics at both the segment level and the system level. In the case of EN→DE, the segments are paragraphs, while in all other cases they are typically sentences. For metrics that use the reference, HE↔EN are scored against refB (a higher-quality reference translation than refA), while EN→DE and ZH→EN are scored against refA.

For EN→DE, HE→EN, and ZH→EN, we have access to human scores for all submitted WMT MT systems (but not for the challenge set systems).

These take the form of MQM scores over a consistent subset of the test set. For the remainder of this work unless otherwise noted, in order to make appropriate comparisons between metric scores on the challenge set, metric scores on the submitted WMT MT systems, and the human annotations, we restrict our analysis to only those segments that are in the news domain and that correspond to the set for which we have human annotations (104 paragraphs for EN→DE, 619 segments we use all 516 segments of the test set data that are in the news domain because we do not have any human annotations).

6.1 Segment Level

Since we only have human annotations for the WMT MT submissions and not our challenge set, we must be cautious in the conclusions that we draw about metric *performance* from the scores they assign to segments. However, we can observe that different metrics exhibit different *characteristics*, even as they score an identical set of segments over an identical set of systems.

6.1.1 Distributions of Scores

As we see in the histograms along the diagonal of Figure 1, showing a subset of the baseline and submitted metrics, different metrics exhibit very different score distributions. This can also be seen in Figures 3, 4, 5, and 6 in Appendix G. Some show a somewhat bimodal distribution of scores, some are closer to normally distributed, and there are a number of metrics whose score distributions do not fall into either of those patterns. Additionally, they differ in whether they exhibit a strong separation between the segments produced by the low-quality systems from our challenge set and the segments produced by the WMT submissions or whether they assign a range of low to high scores to most systems (i.e., having clear overlap in score range across all systems). While we cannot conclude that any of these metrics is more *accurate*, we can note that their varied characteristics suggest that they may be measuring different things and/or that different metrics may have different strengths and weaknesses across the translation quality landscape.

There are also metrics that use an approximation of a discrete score space, such as GEMBA-MQM. This particular metric also scores nearly all segments produced by our low-quality systems as the lowest available score, particularly for EN→DE, meaning it would not be a suitable metric to distinguish between low-to-mid quality (e.g., low-resource) translation systems. XCOMET-Ensemble assigns a wider range of scores to the low-quality segments, but the range and distribution of those scores is fairly consistent across the low-quality systems in our challenge set, meaning that it also struggles to distinguish between system quality levels at the lower end, albeit for a different reason. We can also see this when we examine system-level scores.

6.1.2 Universal Translations and Universal Scores

In Yan et al. (2023), the authors observe what they term "universal translations": target language output that receives high scores regardless of the reference to which they are compared. Here, we observe what one might consider to be "universal scores" instead. Some metrics, like Calibri-COMET22, use a wide range of scores in general, but have a very small subset of scores that appear a very large number of times. For the set of annotated news segments across all challenge set and WMT MT systems for EN→DE, 1673 unique scores are assigned to segments. The vast majority occur only once, but there are two non-minimum/maximum scores that occur 210 and 206 times, respectively (the score zero, i.e. the maximum score for perfect translation in this case, also appears 206 times). In contrast, COMET assigns 2446 unique scores over the same subset of segments, with the most frequent of those scores occurring 7 times. We note that Calibri-COMET22 (and Calibri-COMET22-QE) exhibit this frequently-appearing-score characteristic across the different language pairs, though the number and exact value of the extremely frequent scores differs across language pairs. Importantly, this is not explainable by the data itself: other metrics assign a wide variety of scores to the same segments that receive these particularly common scores, which makes the common scores visible in the Calibri-COMET22 column as the apparent vertical lines (most visible in comparison to COMET). As is evident from both the histogram and the scatterplots, these common scores are most frequently assigned to the segments in our challenge set, to the extent that this unexpected characteristic is not clearly visible when the plot is restricted to only the WMT MT submissions rather than including the challenge set (see Figure 10). This highlights the importance of performing evaluation over a wide

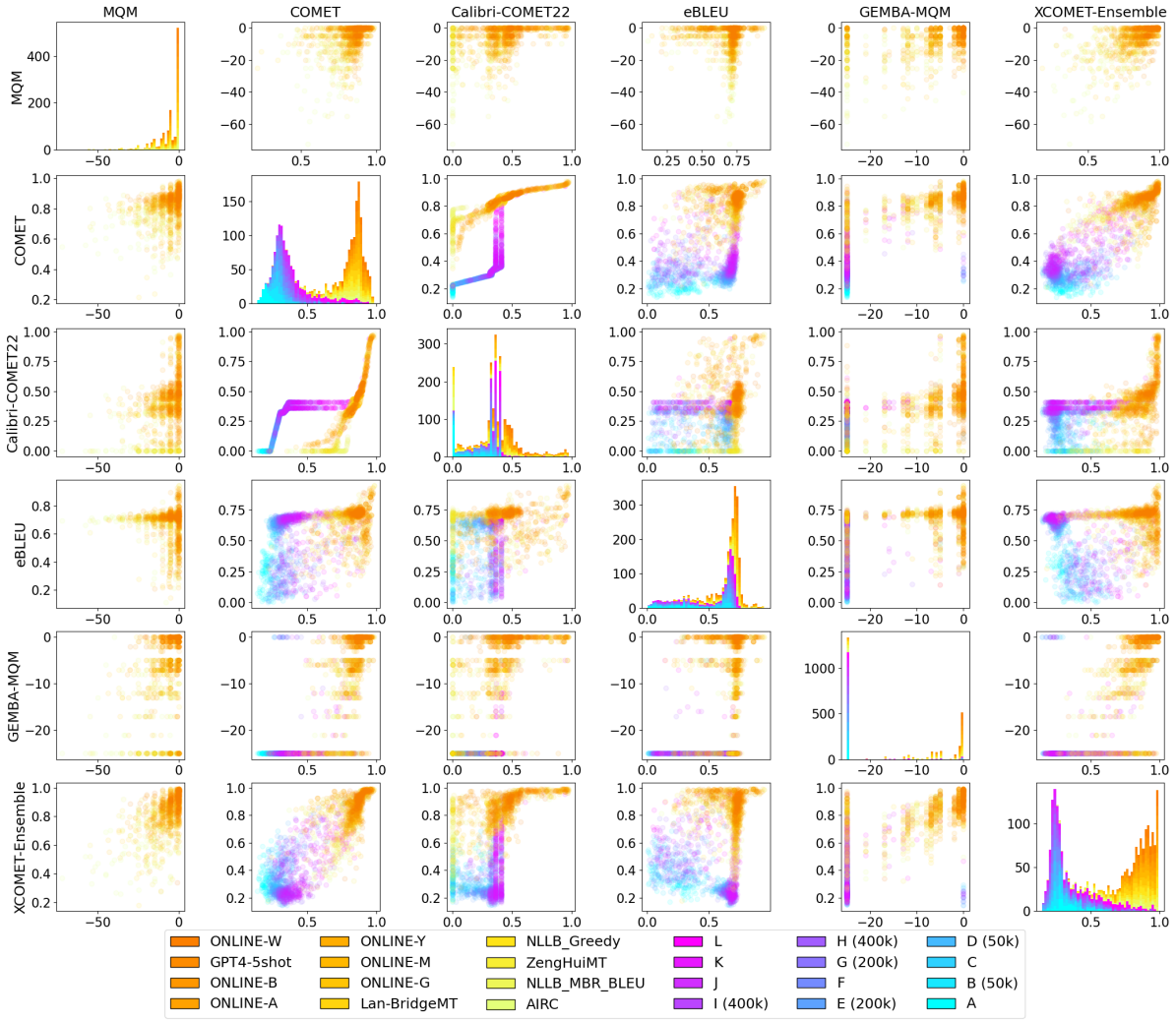

Figure 1: A subset of the metrics (and MQM scores) for EN→DE. The diagonal entries show stacked histograms of segment scores across the challenge set (cool colours/bottom) and submitted WMT systems (warm colours/top). The off-diagonal entries are scatterplots where each point is a single segment positioned according to the score assigned to it by row and column metrics; each point is coloured according to the MT system that produced it.

range of MT quality, in order to discover unexpected issues like this prior to applying the metrics to low-resource or otherwise low-quality MT.

The eBLEU metric exhibits a slightly more dispersed version of this, where a large number of segments receive scores in a fairly narrow band relative to the metric's overall score distributions. However, in the case of eBLEU, this is not specific to the challenge set data, but is also observed in the WMT MT data.

6.2 System Level

To analyze system-level scores, we produce them as an average over all of the segment-level scores in the restricted test set (news domain segments in the set of segments for which WMT MT systems were human-annotated) for a given MT system.[3] These system-level scores can also be used in order to gain a better understanding of the overall range of a metric's scores, as well as what kind of scores are assigned to very low quality machine translation (e.g., the A and B systems from the challenge set).

In Figure 2 (as well as Figures 7, 8, and 9 in Appendix G), we observe that metrics show different patterns of scores at the system level. We observe that some metrics exhibit unexpected characteristics on the low-quality data, such as MaTESe, which ranks some of the low-quality systems in reverse order.[4]

We do not have MQM scores for any of the data in the challenge set, which means that we do not know how much of a gap in quality there is between our best low-quality system and the lowest-performing MT systems submitted to WMT. However, we can observe that metrics differ widely in their estimates of the gap; embed_llama, for example, shows error bar overlap between the highest performing system from our challenge set and the lowest-scored system from the submissions, while GEMBA-MQM shows a very large gap between the two groups of systems,[5] with many of the other metrics falling between these extremes.

Similarly, again examining characteristics without making claims about metric performance, we notice variety amongst the metrics in terms of the range of scores they assigned to each group of systems, as seen in the slope of the system scores. In some cases, there are quite similar slopes (e.g., embed_llama), while in other cases there is a steep slope for the challenge set as compared to the WMT MT systems (e.g., BERTScore or COMET) meaning that the challenge set covers a wide range of (lower) scores while the WMT MT set covers a smaller range of higher scores, and finally some systems where the slopes are similar but both less steep (e.g., Calibri-COMET22-QE and GEMBA-MQM) and each set of systems covers a small range of scores with a gap in between. Without MQM scores for the challenge set, we do not know whether one of these patterns is indicative of a metric that more closely resembles human annotations or not (i.e., we do not know whether the challenge set covers a wider range of quality than the WMT MT systems, which would support metrics having a steeper slope/wider range in the scores assigned to it).

We also note some variety across language pairs. The reversal of scores seen in MaTESe is less obvious in the EN→DE data, though that may be related to greater overlap in the EN→DE challenge set data quality.

In future work, obtaining MQM scores for one or more of the systems in our challenge set would permit us to draw conclusions about metric performance in these areas (i.e., about whether there is indeed quality overlap between the two sets, and what appropriate ranges of scores might be for each of the sets).

All of these observations about variation in metric characteristics raise an important issue in the evaluation and adoption of new metrics: since their correlation with human rankings is often demonstrated on the high-quality MT output being scored at WMT, it is not necessarily appropriate to use them for the evaluation of low-resource or lower-quality MT output without additional study.

6.3 Additional Discussion

We briefly mention two other items of note from our exploration of the data.

Outside of the set of data for which there are human annotations (i.e., not appearing in our figures), for HE→EN there are 14 news domain segments for which the ZengHuiMT system output an empty string. Different metrics handle this in

[3] This includes the baseline BLEU.

[4] Though we do not have extensive human evaluation, we are confident that, e.g., system E should not be ranked below system A.

[5] For EN→DE, in Figure 7, the Calibri-COMET22 metrics both find several of the highest performing systems from our challenge set to be better than several of the submitted systems, while most other metrics rank the challenge set systems below the submitted systems.

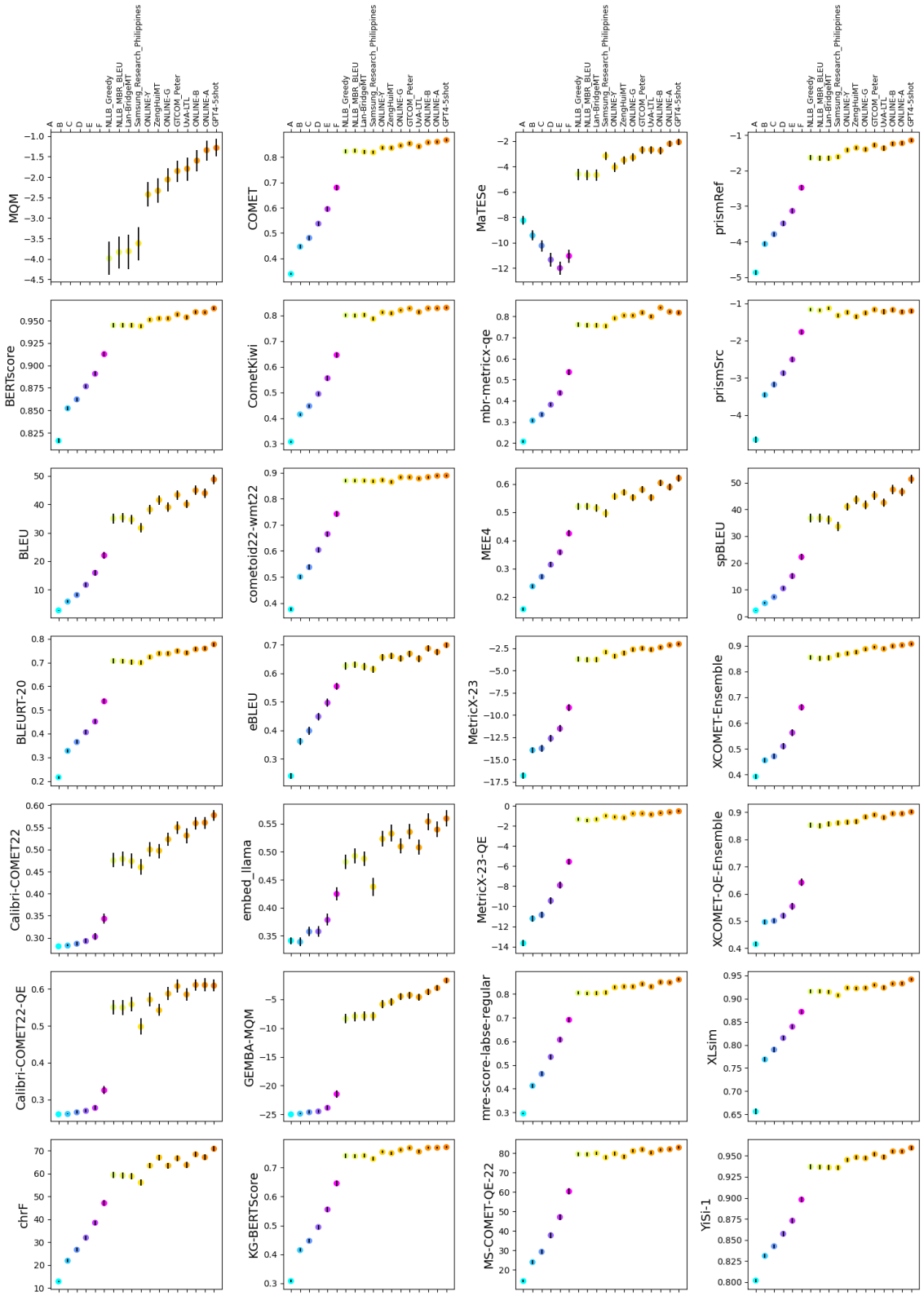

Figure 2: System average scores (with error bars computed via bootstrap resampling 1000 times for $p < 0.05$) for HE→EN across the challenge set (cool colours/left) and submitted WMT systems (warm colours/right). Our challenge set systems are ordered from left to right with BLEU scores, while the submitted WMT systems are ordered by MQM score on the news domain.

Metric	Score	Range
BERTscore	0.000	(0.000, 1.000)
BLEU	0.000	(0.000, 100.000)
BLEURT-20	0.055	(0.000, 1.030)
Calibri-COMET22	0.328	(0.000, 0.990)
Calibri-COMET22-QE	0.083	(0.000, 1.000)
chrF	0.000	(0.000, 100.000)
COMET*	0.796	(0.287, 0.995)
CometKiwi*	0.647	(0.261, 0.902)
cometoid22-wmt22	0.597	(0.268, 0.994)
eBLEU	0.000	(0.000, 1.000)
embed_llama	0.510	(0.040, 1.000)
GEMBA-MQM	-25.000	(-25.000, 0.000)
KG-BERTScore*	0.682	(0.285, 0.886)
MaTESe	0.000	(-25.000, 0.000)
mbr-metricx-qe	0.027	(-0.004, 0.998)
MEE4	0.000	(0.000, 1.000)
MetricX-23*	-25.597	(-25.618, 0.198)
MetricX-23-QE*	-24.546	(-24.557, 0.848)
mre-score-labse-regular*	0.772	(0.266, 0.965)
MS-COMET-QE-22*	59.243	(1.641, 94.075)
prismRef	-5.256	(-8.685, -0.077)
prismSrc	-6.829	(-10.027, -0.111)
spBLEU	0.000	(0.000, 100.000)
XCOMET-Ensemble*	0.917	(0.291, 0.994)
XCOMET-QE-Ensemble*	0.899	(0.290, 0.998)
XLsim	0.911	(0.569, 1.000)
YiSi-1	0.000	(0.000, 1.000)

Table 1: Average metric scores assigned to empty strings in the HE→EN news data, shown with the full range of metric scores assigned to the news data. Metrics with asterisks by their name did not assign the same scores to all the empty strings, though the differences were quite small.

different ways; some assign a score of 0 (or the metric's lower bound score), while others assigned relatively high scores due to the fact that the source and reference were very short (each source and reference consisted only of a single period). Table 1 shows the scores assigned by metrics to these empty strings as well as the range of scores over the full HE→EN news data (including segments that were not human-annotated, as the empty strings were also not included for human annotation).

We also observe two examples of systems that receive noticeably lower scores from a number of metrics than would be expected based on their human ranking: NLLB_Greedy (EN→DE, Figure 3) and Samsung_Research_Philippines (HE→EN, Figure 5). We leave this as an area for future investigation.

7 Conclusions

This challenge set expands the range of system quality scored by metrics at the shared task. This expanded range of MT quality reveals interesting characteristics and limitations of some new metrics when applied to a broader range of systems. The smaller variations in segment-level scores given by some metrics at the low end of quality could indicate that these metrics struggle to discriminate low-quality MT systems. This is further shown by the observation that some metrics rank the low-quality systems in reverse order at the system level. We have discovered a "universal score" phenomenon for some metrics, where a small subset of non-minimum/maximum distinct scores are assigned to a variety of translation output. This characteristic was not visible in the high-quality MT output alone, highlighting the importance of this type of testing. We also observe diverse behaviors from different metrics on empty string translation.

Our challenge set serves as a complement to the standard correlation-based analyses and also provides useful information to researchers who are considering using these metrics in low-resource or low-quality domains. We recommend that metric researchers check their metrics' performance on a wider landscape of translation quality or be clear about the limitations of their metrics' testing.

Limitations

A major limitation of this work is our choice to select low-quality systems on the basis of BLEU scores, which was done for reasons of time and cost. We attempted to mitigate this by spot-checking to confirm that we saw noticeable differences between various pairs of low-quality systems, but a more thorough human annotation would be beneficial. We are also limited in the set of languages we have explored, using only four language pairs, as well as the limited news domain.

Acknowledgements

We thank Adam Poliak for comments and feedback on a sample of the English–Hebrew dataset. We thank our colleagues and the anonymous reviewers for feedback and suggestions as well as the WMT Metrics Task and General Task organizers for providing data and annotations.

References

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu

Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eleftherios Avramidis and Vivien Macketanz. 2022. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 514–529, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 346–356, Online. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017a. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017b. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, and Ying Qin. 2022. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Modifying kendall's tau for modern metric meta-evaluation.

Sören DREANO, Derek Molloy, and Noel Murphy. 2023. Embed_Llama: using LLM embeddings for the Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad ElNokrashy and Tom Kocmi. 2023. eBLEU: Using Simple Word Embeddings For Efficient Machine Translation Evaluation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

M. Amin Farajian, Rajen Chatterjee, Costanza Conforti, Shahab Jalalvand, Vevake Balaraman, Mattia A. Di Gangi, Duygu Ataman, Marco Turchi, Matteo Negri, and Marcello Federico. 2016. FBK's neural machine translation systems for IWSLT 2016. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Thamme Gowda, Tom Kocmi, and Marcin Junczys-Dowmunt. 2023. Cometoid: Distilling Strong Reference-based Machine Translation Metrics into Even Stronger Quality Estimation Metrics. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality

estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden. Association for Computational Linguistics.

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

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. 2022. Sockeye 3: Fast neural machine translation with pytorch.

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 (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Christof Monz, Makoto Morishita, Murray Kenton, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting Quality Error Spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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.

Chi-kiu Lo, Rebecca Knowles, and Cyril Goutte. 2023. Beyond correlation: Making sense of the score differences of new mt evaluation metrics. In *Proceedings of Machine Translation Summit XIX Vol. 1: Research Track*, pages 186–199.

Vivien Macketanz, Renlong Ai, Aljoscha Burchardt, and Hans Uszkoreit. 2018. TQ-AutoTest – an automated test suite for (machine) translation quality. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Benjamin Marie. 2022. Yes, we need statistical significance testing. https://pub.towardsai.net/yes-we-need-statistical-significance-testing-927a8d21f9f0.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Toronto, Canada. Association for Computational Linguistics.

Anthony Moi and Nicolas Patry. 2022. Huggingface's tokenizers.

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.

Subhajit Naskar, Daniel Deutsch, and Markus Freitag. 2023. Quality Estimation using Minimum Bayes Risk. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Mark Przybocki, Kay Peterson, Sébastien Bronsart, and Gregory Sanders. 2009. The nist 2008 metrics for machine translation challenge—overview, methodology, metrics, and results. *Machine Translation*, 23(2/3):71–103.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

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

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.

Vasiliy Viskov, George Kokush, Daniil Larionov, Steffen Eger, and Alexander Panchenko. 2023. Semantically-Informed Regressive Encoder Score. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Zhanglin Wu, Yilun Liu, Min Zhang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Xiaosong Qiao, Jingfei Zhang, Ma Miaomiao, Zhao Yanqing, Song Peng, Shimin Tao, Hao Yang, and Yanfei Jiang. 2023. Empowering a Metric with LLM-assisted Named Entity Annotation: HW-TSC's Submission to the WMT23 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Corpora Sizes

Corpora are from WMT23 (Kocmi et al., 2023) and WMT17 (Bojar et al., 2017a).

A.1 EN→DE

From the download table at EMNLP 2023: General Machine Translation, we retrieved all **EN→DE** corpora. The *train* corpus is composed of *airbaltic*, *czechtourism*, *ecb2017*, *EESC2017*, *EMA2016*, *rapid2016*, *europarl-v10*, *news-commentary-v18*, *WikiMatrix.v1.de-en.langid* and *wikititles-v3*. We chose *newstest2020* as our *validation*. Corpora statistics are described in Table 2.

Name	# lines	# de words	# en words
train	14,227,278	234,635,104	246,351,534
validation	1418	45,855	44,018

Table 2: Corpora sizes for EN→DE, computed on raw text (not tokenized) using `wc`.

A.2 HE↔EN

From the instructions of EMNLP 2023: General Machine Translation, we retrieved all **HE→EN** corpora.[6] Then, using *wmt23-heen/train.{heb,eng}* and *WikiMatrix.en-he*, we sampled sentence pairs for the *validation* and *test* sets and the remaining pairs were used for *train* making sure that all three are mutually exclusive. Corpora statistics are described in Table 3.

Name	# lines	# en words	# he words
train	2,227,830	38,307,579	30,943,929
validation	2000	20,459	16,620

Table 3: Corpora sizes for EN→HE, computed on raw text (not tokenized) using `wc`.

A.3 ZH→EN

We used corpora from WMT2017.[7] *train* is composed of all 20 *Books*, *casia2015*, *casict2015*, *casict-A*, *casict-B*, *datum*, *NEU*, *news-commentary-v18.en-zh*, *WikiMatrix.v1.en-zh.langid* and *wikititles-v3*. We chose *newstest2020* as our *validation*. Corpora statistics are described in Table 4.

Name	# lines	# en words	# zh words
train	12,995,613	218,659,998	43,676,661
validation	2000	65,561	3716

Table 4: Corpora sizes for ZH→EN, computed on raw text (not tokenized) using `wc`.

B Subword Segmentation

After some light normalization consisting of converting non-breaking hyphen, normalizing spaces, replacing control characters with spaces and collapsing multiple spaces,[8] we trained a 32k tokens, bilingual sentencepiece unigram subtokenizer using HuggingFace's tokenizers (Moi and Patry, 2022) for each language pair. The corpora used for training the subword model were:

- **EN→HE** uses all of *wmt23-heen/train.{eng,eng}*
- **EN→DE** uses our concatenated *train*
- **ZH→EN** uses our concatenated *train*

C System Descriptions

For all systems, we used Sockeye-3.1.31 (Hieber et al., 2022), commit `13c63be5` with PyTorch-1.12.1 (Paszke et al., 2019). Training was performed on 4 Tesla V100-SXM2-32GB GPUs for **EN→DE** and **ZH→EN** and 4 Tesla V100-SXM2-16GB GPUs for **EN→HE** and **HE→EN**. Training times are shown in Table 5.

Name	Time (h)
ende	6 - 35
enhe	10.3
heen	6.5
zhen	83.5

Table 5: Training times in hours.

Table 6 describes the differences with Sockeye's default parameters. Note that we kept all intermediate checkpoints (from which we later select the outputs used for the challenge set) and used the entire validation during checkpoint evaluation.

[6] `mtdata get-recipe -ri wmt23-heen -o wmt23-heen`

[7] https://www.statmt.org/wmt17/translation-task.html

[8] https://github.com/nrc-cnrc/PortageTextProcessing/blob/main/bin/clean-utf8-text.pl

Name	Value
amp	*True*
grading clipping type	*abs*
max sequence length	*200:200*
batch type	*max-word*
checkpoint interval	*10*
initial learning rate	*0.06325*
learning rate scheduler type	*inv-sqrt-decay*
learning rate warmup	*4000*
max checkpoints	*110*
max epochs	*1000*
max num checkpoint not improved	*32*
optimizer	*Adam*
optimizer Betas	*0.9, 0.98*
optimized metric	*BLEU*
update interval	*10*
attention heads	*16:16*
shared vocabulary	*True*
transformer FFN	*4096:4096*
transformer model size	*1024:1024*
weight tying	*True*

Table 6: Differences from Sockeye's default parameters.

On top of the changes from Table 6, for **EN→HE** and **HE→EN**, we lowered the **batch size** to *6144* and changed **max checkpoints** to *330*.

For all language pairs, we have trained a baseline system using the entire *train* corpus. Additionally, for **EN→DE**, we also trained systems that use a uniformly random subsample of *train* namely, 50k, 200k and 400k (the pseudo-low-resource systems).

D Pseudo-Low-Resource Corpora

Due to human error in the sampling code, the pseudo-low-resource training data used for the EN→DE systems trained on 50k, 200k, and 400k—intended to be a random sample from the full training data—instead primarily consists of data from the first four corpora shown in Table 7. Table 8 shows the small number of differences between these subsampled corpora and simply selecting the first n lines of the full training corpus. The main consequence of this is that these systems may be skewed towards particular domains.

E Checkpoints in Challenge Set

In Table 9 we see the checkpoint IDs for systems included in the challenge set for HE↔EN and ZH→EN. Table 10 shows the same for EN→DE. The corresponding BLEU scores are shown in Tables 11 and 12, respectively.

Corpus Name	# Sentences
airbaltic	839
czechtourism	6758
ecb2017	4147
EESC2017	2,857,850
EMA2016	347,631
rapid2016	1,030,808
europarl-v10	828,473
news-commentary-v18	203,744
WikiMatrix.v1	2,579,106
wikititles	1,474,203
total	14,227,278

Table 7: (EN→DE) Sub-corpora sizes in the order they were merged to create the final sampled *train*.

Sample Size	# Differences	# lines *EESC2017*
50k	282	38,256
200k	70	188,256
400k	34	388,256

Table 8: EN→DE; Number of sentences that are different from the original train's head and how many sentences from *EESC2017* that were used.

System	EN→HE	HE→EN	ZH→EN
A	68	58	61
B	98	87	91
C	115	102	115
D	135	117	139
E	171	140	222
F	392	219	480

Table 9: Checkpoint IDs for systems included in challenge set (HE↔EN and ZH→EN).

System	EN→DE
A	54
B (50k)	1
C	79
D (50k)	7
E (200k)	2
F	91
G (200k)	27
H (400k)	4
I (400k)	43
J	102
K	129
L	313

Table 10: Checkpoint IDs for systems included in challenge set (EN→DE); parenthetical numbers indicate one of the pseudo-low-resource systems rather than the full training data system.

System	EN→HE	HE→EN	ZH→EN
A	0.6	0.7	0.9
B	3.1	4.3	5.0
C	7.2	7.3	9.3
D	11.4	11.4	13.1
E	16.6	16.0	18.5
F	26.2	23.9	23.2

Table 11: BLEU scores for systems included in challenge set over the full news data in the challenge set (HE↔EN computed with refB).

System	EN→DE
A	0.7
B (50k)	2.5
C	4.2
D (50k)	4.7
E (200k)	4.7
F	8.9
G (200k)	9.5
H (400k)	10.4
I (400k)	12.0
J	12.8
K	18.7
L	29.9

Table 12: BLEU scores for systems included in challenge set (EN→DE) over the full news data in the challenge set; parenthetical numbers indicate one of the pseudo-low-resource systems rather than the full training data system.

F Metrics

Table 13 shows a summary of the human annotations and metrics included in this work and the translation directions they participated in. In the following, we briefly describe the key characteristic of each metric.

F.1 Baseline Metrics

BLEU (Papineni et al., 2002) is the (clipped) precision of word n-grams between the MT output and its reference weighted by a brevity penalty.

spBLEU (Team et al., 2022) is BLEU computed with subword tokenization done by the Flores-200 Sentencepiece Model (Kudo and Richardson, 2018).

chrF (Popović, 2015) uses character n-grams to compare the MT output with the reference and it is a balance of precision and recall.

BERTScore (Zhang et al., 2020) uses cosine similarity of contextual embeddings from pre-trained transformers to compute F-scores of sentence level similarity.

BLEURT-20 (Sellam et al., 2020) is fine-tuning RemBERT to predict direct assessment (DA; Graham et al., 2013, 2014, 2016) scores for a MT-reference pair.

COMET (COMET-22) (Rei et al., 2022) is an ensemble of two models: COMET-20 and a multitask model jointly predicting sentence-level multidimensional quality metrics (MQM) and word-level translation quality annotation, where COMET-20 is fine-tuning XLM-R to predict DA scores for a MT-source-reference tuple. **CometKiwi** is a quality estimation metric that is similar to COMET, except it scores the MT output against the source, instead of the reference translation.

MS-COMET-QE-22 (Kocmi et al., 2022) is a COMET-QE-20 based quality estimation metric trained on a larger and filtered set of human judgements, covering 113 languages and 15 domains.

prismRef (Thompson and Post, 2020) uses a neural paraphrase model to score the MT output against the reference translation. **prismSrc** is the quality estimation version, which scores the MT output against the source, instead of the reference translation.

YiSi-1 (Lo, 2019) measures the semantic similarity between the MT output and reference by the IDF-weighted cosine similarity of contextual embeddings extracted from pretrained language models, e.g. RoBERTa, CamemBERT, XLM-R, etc., depending on the target language in evaluation.

F.2 Primary submissions

Calibri-COMET22 uses isotonic regression on the COMET-22 output scores to predict the fraction of translations with no error produced by the MT system. **Calibri-COMET22-QE** is a quality estimation metric that is similar to Calibri-COMET22, where it uses COMETKiwi as base model.

cometoid22-wmt22 (Gowda et al., 2023) is a quality estimation metric that uses COMET-22 as a teacher metric and trains a student model to predict the teacher scores without using reference translation.

eBLEU (ElNokrashy and Kocmi, 2023) uses non-contextual word embeddings and relative meaning diffusion tensors to approximate the token similarity in the MT output and reference and computes translation quality scores similar to BLEU.

embed_llama (DREANO et al., 2023) is the

Metric Name	EN→DE	EN→HE	HE→EN	ZH→EN	Reference-based
Human annotation					
MQM	✓		✓	✓	
Metrics					
BERTScore	✓	✓	✓	✓	✓
BLEU	✓	✓	✓	✓	✓
BLEURT-20	✓	✓	✓	✓	✓
Calibri-COMET22	✓	✓	✓	✓	✓
Calibri-COMET22-QE	✓	✓	✓	✓	
chrF	✓	✓	✓	✓	✓
COMET	✓	✓	✓	✓	✓
CometKiwi	✓	✓	✓	✓	
cometoid22-wmt22	✓	✓	✓	✓	
eBLEU	✓	✓	✓	✓	✓
embed_llama	✓	✓	✓	✓	✓
GEMBA-MQM	✓	✓	✓	✓	
KG-BERTScore	✓	✓	✓	✓	
MaTESe	✓		✓	✓	✓
mbr-metricx-qe	✓		✓	✓	
MEE4	✓	✓	✓	✓	✓
MetricX-23	✓	✓	✓	✓	✓
MetricX-23-QE	✓	✓	✓	✓	
mre-score-labse-regular	✓	✓	✓	✓	✓
MS-COMET-QE-22	✓	✓	✓	✓	
prismRef	✓	✓	✓	✓	✓
prismSrc	✓	✓	✓	✓	
spBLEU (flores-200)	✓	✓	✓	✓	✓
XCOMET-Ensemble	✓	✓	✓	✓	✓
XCOMET-QE-Ensemble	✓	✓	✓	✓	
XLsim	✓	✓	✓	✓	✓
YiSi-1	✓	✓	✓	✓	✓

Table 13: Human annotation and metrics included in this work, with their coverage of language pairs. Metrics that are not marked as reference-based are reference-free (a.k.a quality estimation) metrics.

cosine similarity of the MT output and reference based on Llama 2 sentence embeddings.

GEMBA-MQM (Kocmi and Federmann, 2023) uses three-shot prompting on the GPT-4 model with a single prompt and no language specific example.

KG-BERTScore (Wu et al., 2023) is the linear combination of KGScore and COMET-QE based BERTScore, where KGScore is incorporating multilingual knowledge graph into BERTScore.

MaTESe (Perrella et al., 2022) trains Deberta (for English) and InfoXLM (for German and Russian) encoders to identify MQM error spans and severity using WMT22 Metrics shared task MQM data.

mbr-metricx-qe (Naskar et al., 2023) uses the underlying technique of minimum bayes risks (MBR) decoding to develop a quality estimation metric. It uses an evaluator machine translation system and a reference-based utility metric (specifically BLEURT and MetricX) to calculate a quality estimation score of a model.

MEE4 (Mukherjee and Shrivastava, 2023) is an unsupervised, reference-based metric that is a weighted combination of syntactic similarity based on a modified BLEU score, lexical, morphological and semantic similarity using unigram matching and contextual similarity with sentence similarity scores from multilingual BERT.

MetricX-23 (Juraska et al., 2023) is a regression metric that finetunes the mT5-XXL checkpoint using direct assessment data from 2015-2020 and MQM data from 2020 to 2021 as well as synthetic data. **MetricX-23-QE** is the quality estimation variant that uses the source, instead of the reference, for scoring.

mre-labse-regular (Viskov et al., 2023) is a trained metric that is based on the encoder part of mT0-large model and contextual embeddings from LaBSE. It concatenates the source, reference and MT output as input.

XCOMET-Ensemble (Guerreiro et al., 2023) is an ensemble of a XCOMET-XL and two XCOMET-XXL checkpoints that result from the different training stages. XCOMET is similar to COMET but is trained for both regression and sequence tagging for identifying MQM error spans, where the intent is to make it a more interpretable learnt metric. **XCOMET-QE-Ensemble** is the quality estimation version.

XLsim (Mukherjee and Shrivastava, 2023) is a supervised reference-based metric that regresses on human scores provided by WMT (2017-2022) based on XLM-RoBERTa using a Siamese network architecture with CosineSimilarityLoss.

G Additional Figures

Here we show additional figures, including the full set of histograms for EN→DE (Figure 3), EN→HE (Figure 4), HE→EN (Figure 5) and ZH→EN (Figure 6) as well as the system scores for EN→DE (Figure 7), EN→HE (Figure 8), and ZH→EN (Figure 9).

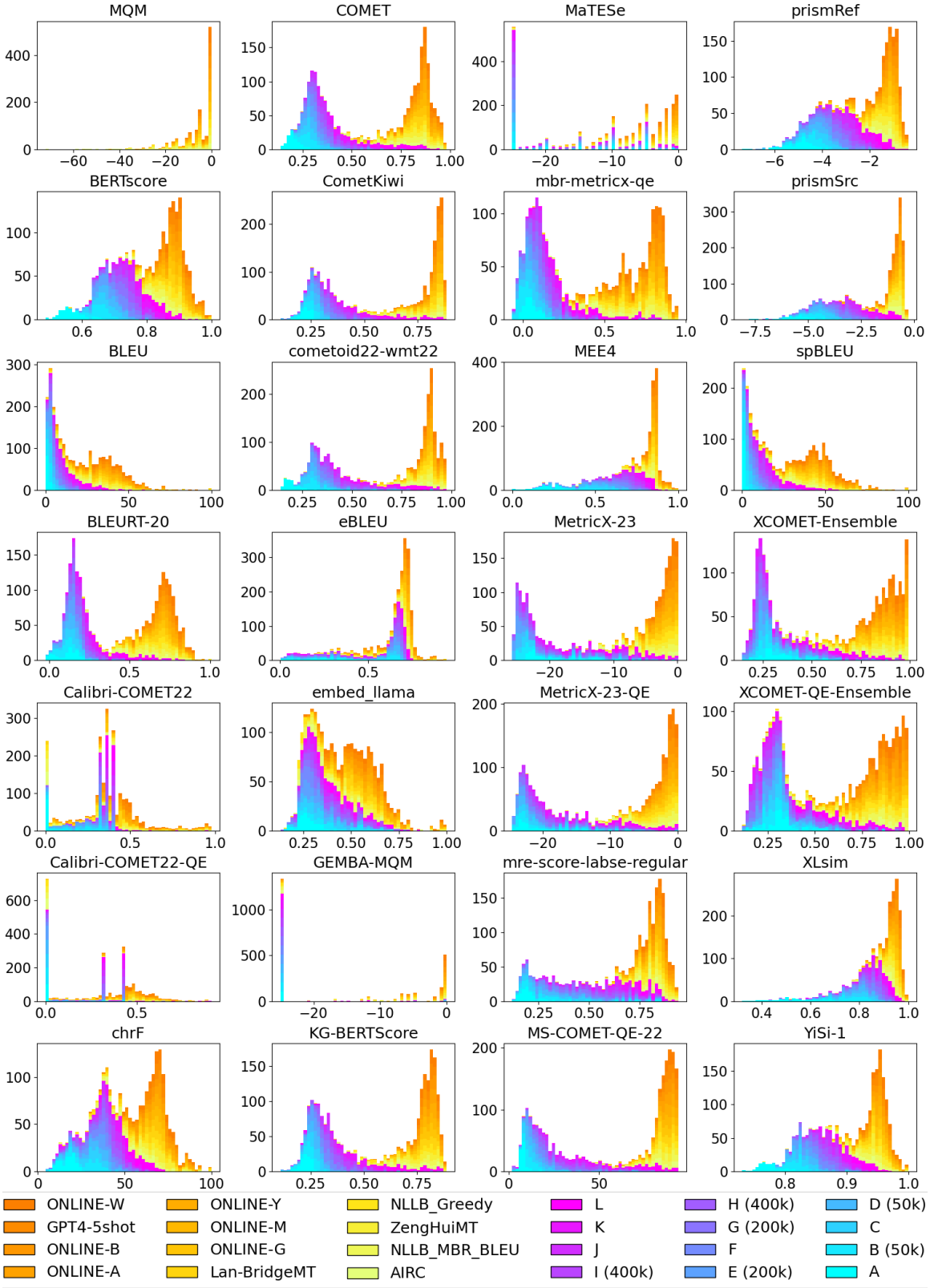

Figure 3: Stacked histograms (one subplot per metric) of segment scores for EN→DE across the challenge set (cool colours/bottom of the stacked histograms) and submitted WMT systems (warm colours/top of the stacked histograms).

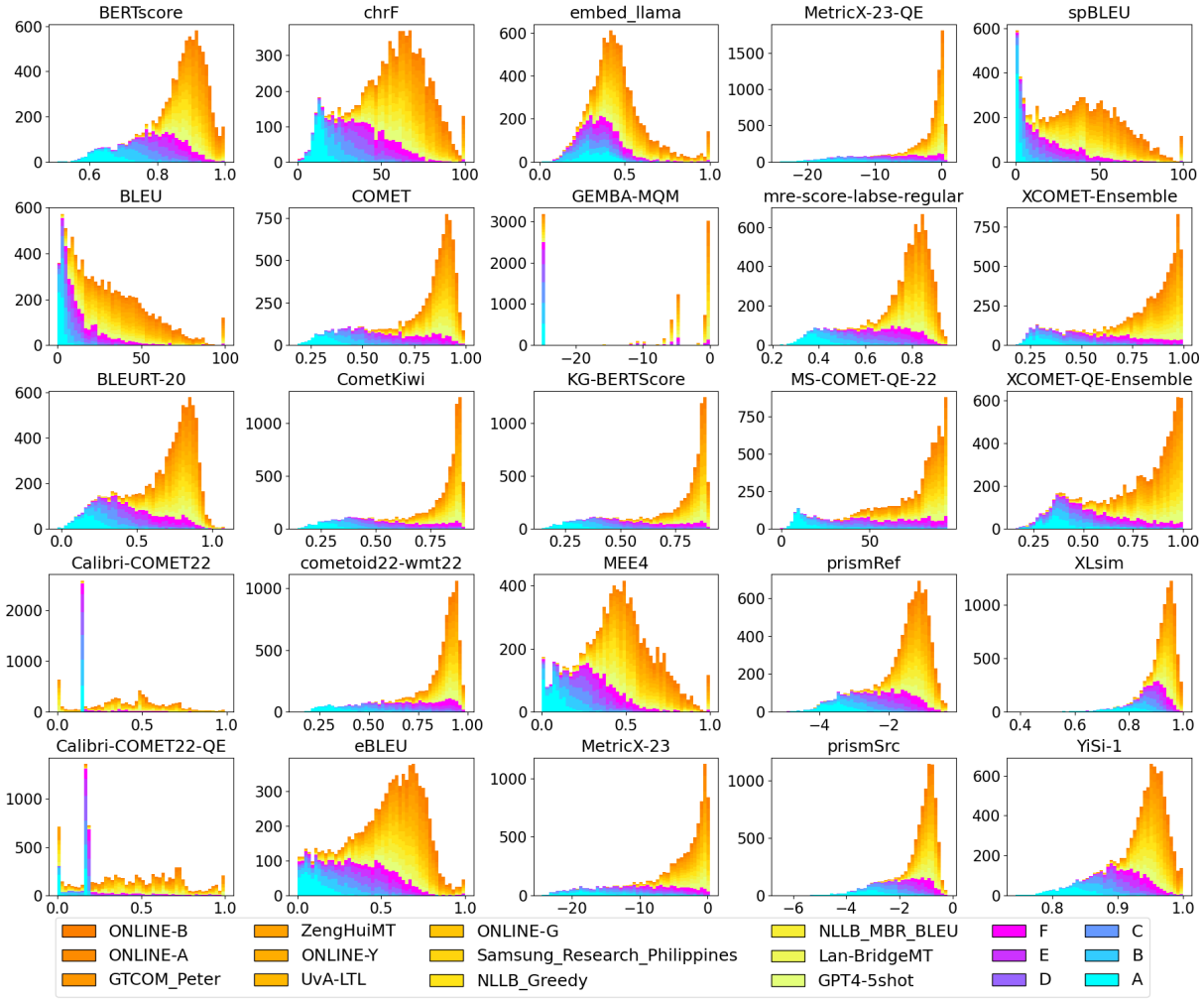

Figure 4: Stacked histograms of segment scores for EN→HE across the challenge set (cool colours/bottom) and submitted WMT systems (warm colours/top).

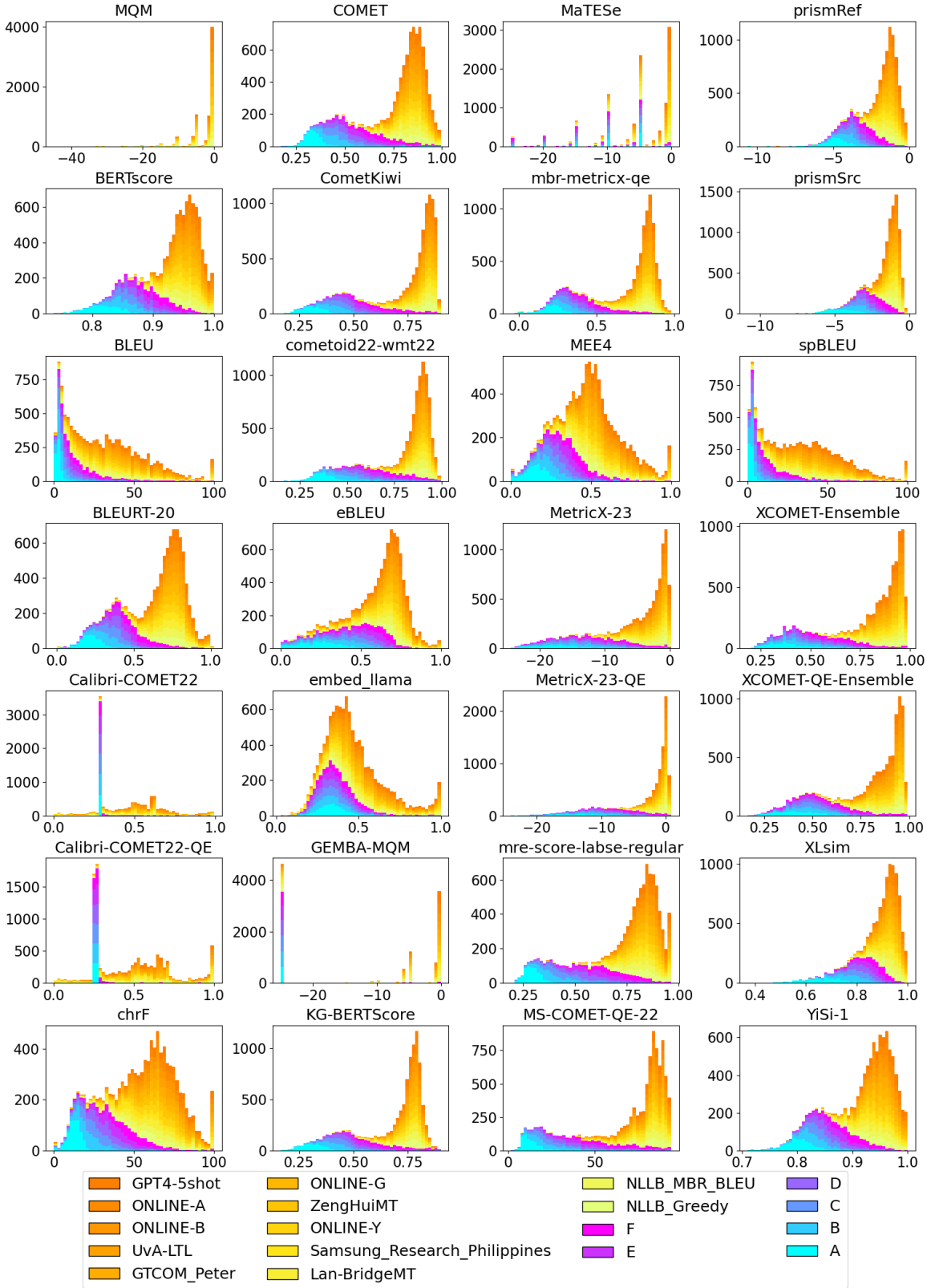

Figure 5: Stacked histograms of segment scores for HE→EN across the challenge set (cool colours/bottom) and submitted WMT systems (warm colours/top).

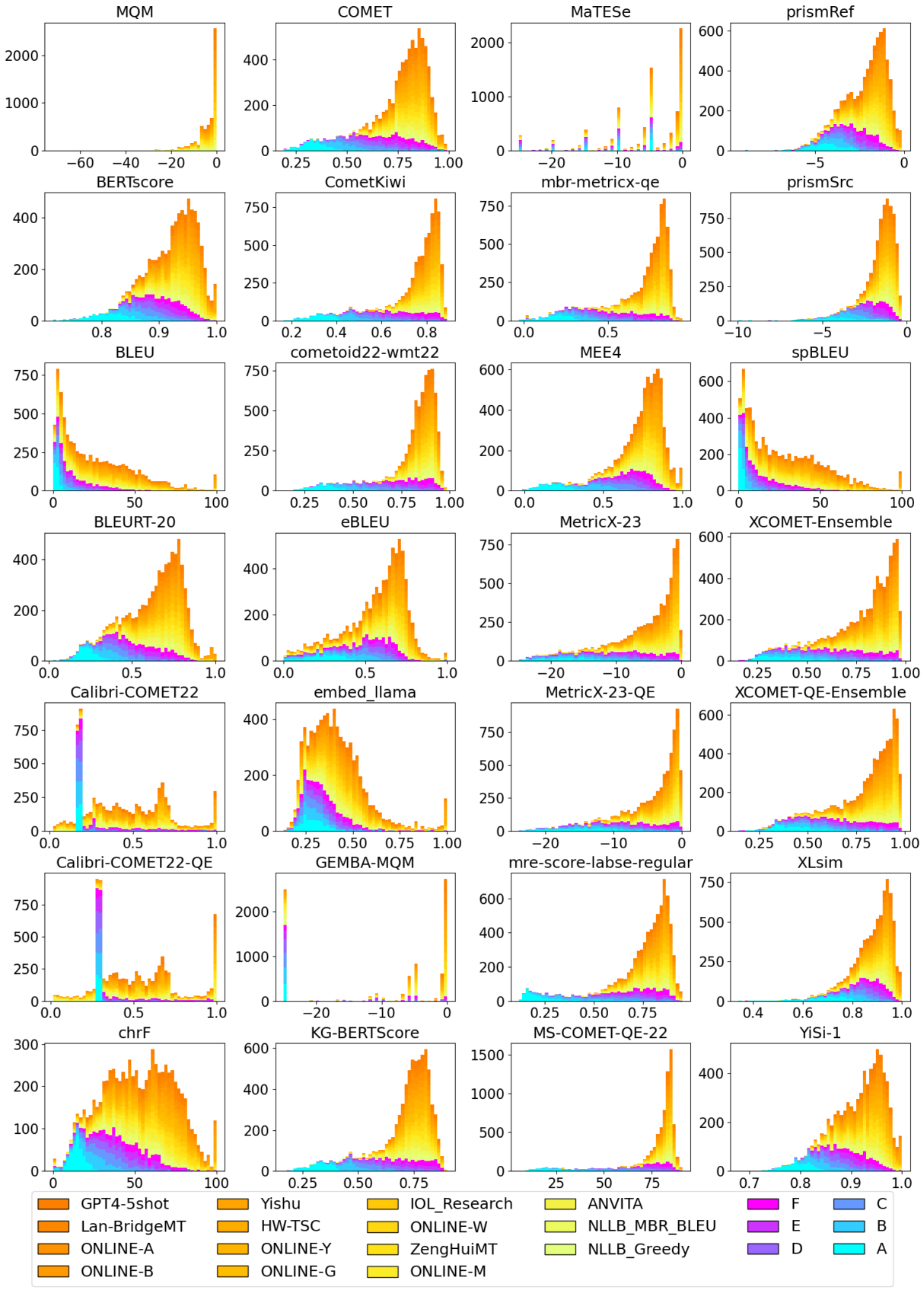

Figure 6: Stacked histograms of segment scores for ZH→EN across the challenge set (cool colours/bottom) and submitted WMT systems (warm colours/top).

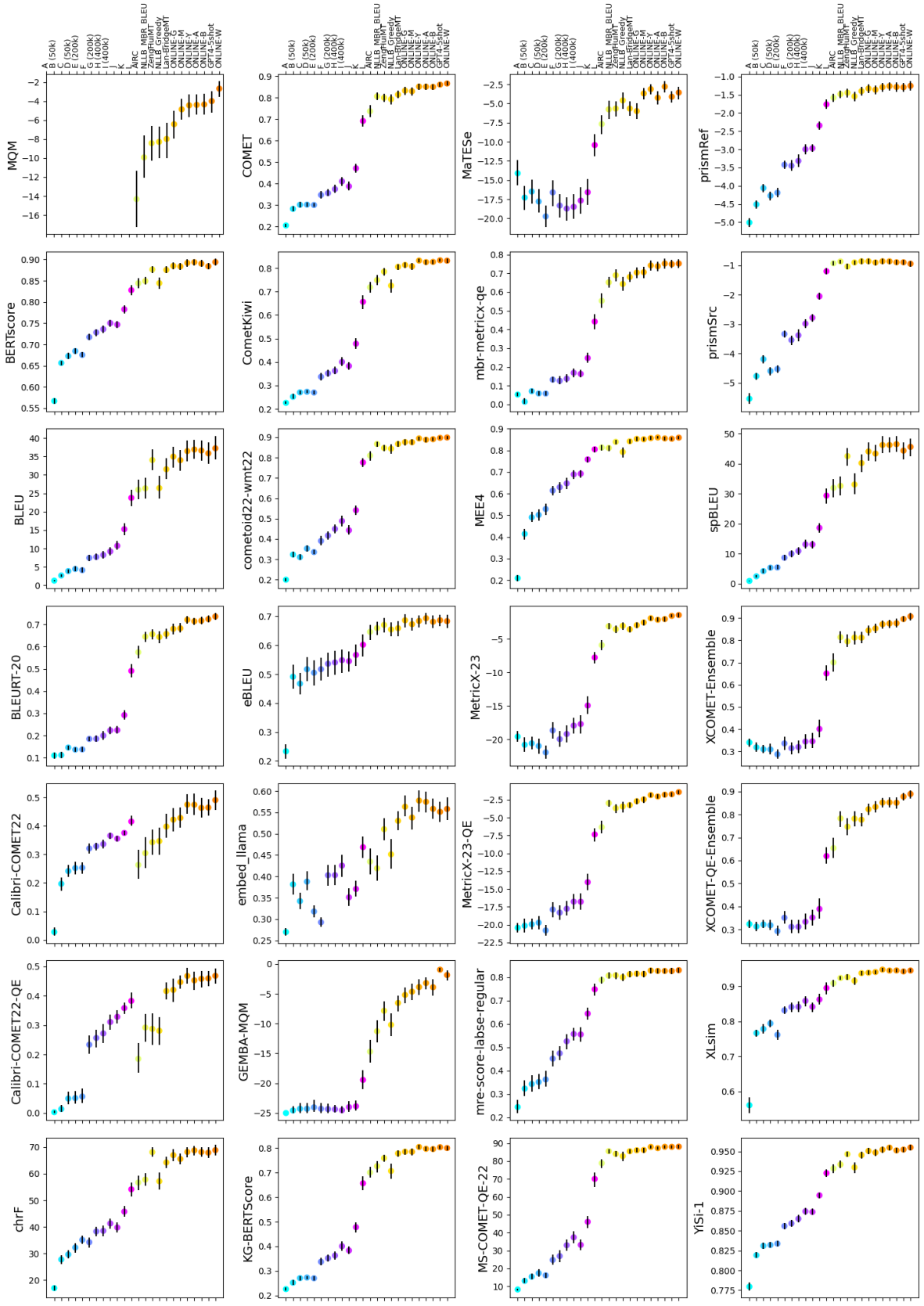

Figure 7: System average scores (with error bars computed via bootstrap resampling 1000 times for $p < 0.05$) for EN→DE across the challenge set (cool colours/left) and submitted WMT systems (warm colours/right). Our challenge set systems are ordered from left to right with BLEU scores, while the submitted WMT systems are ordered by MQM score on the news domain.

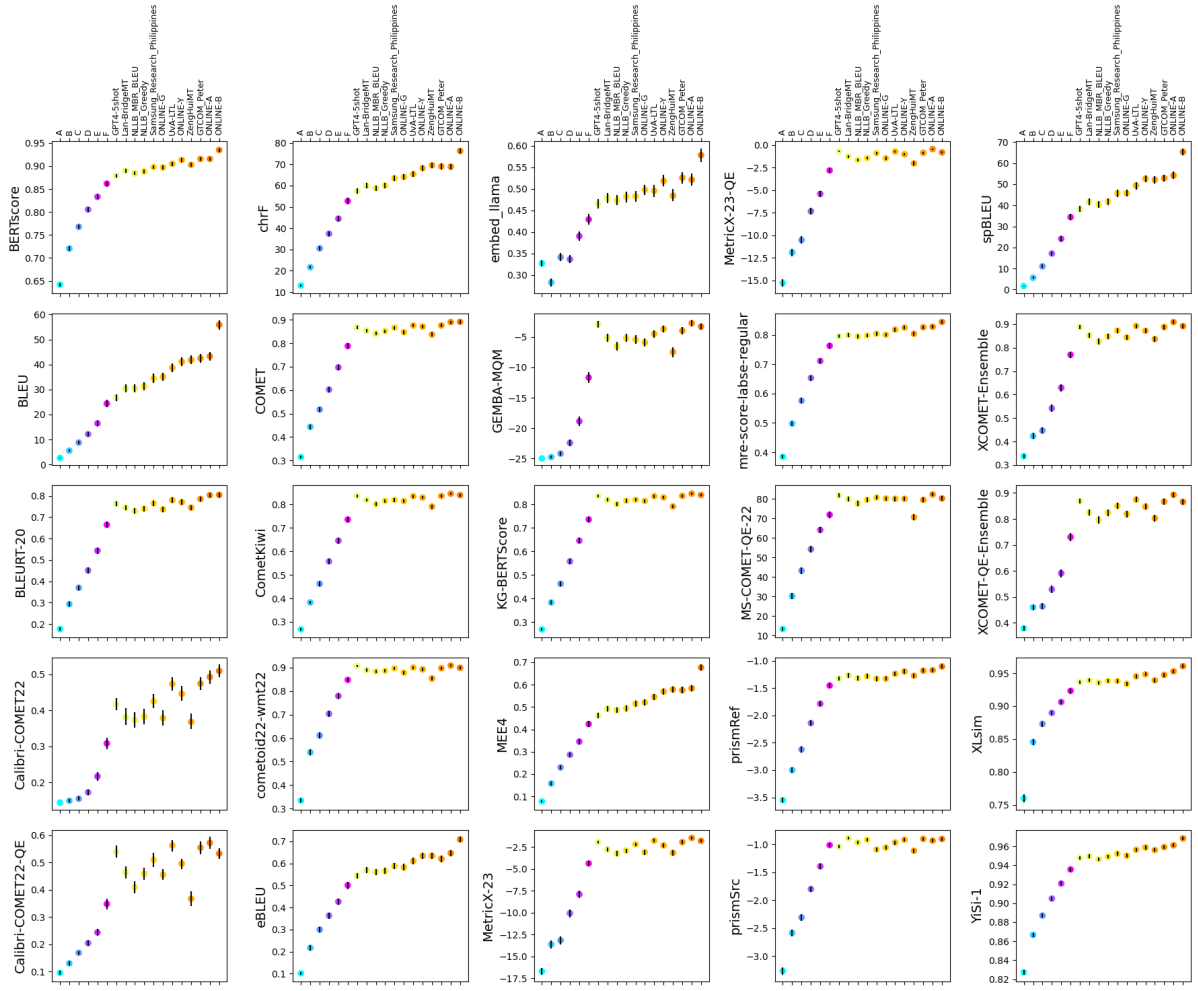

Figure 8: System average scores (with error bars computed via bootstrap resampling 1000 times for $p < 0.05$) for EN→HE across the challenge set (cool colours/left) and submitted WMT systems (warm colours/right). All systems are ordered from left to right by BLEU scores (as direct assessment scores were not yet available for EN→HE).

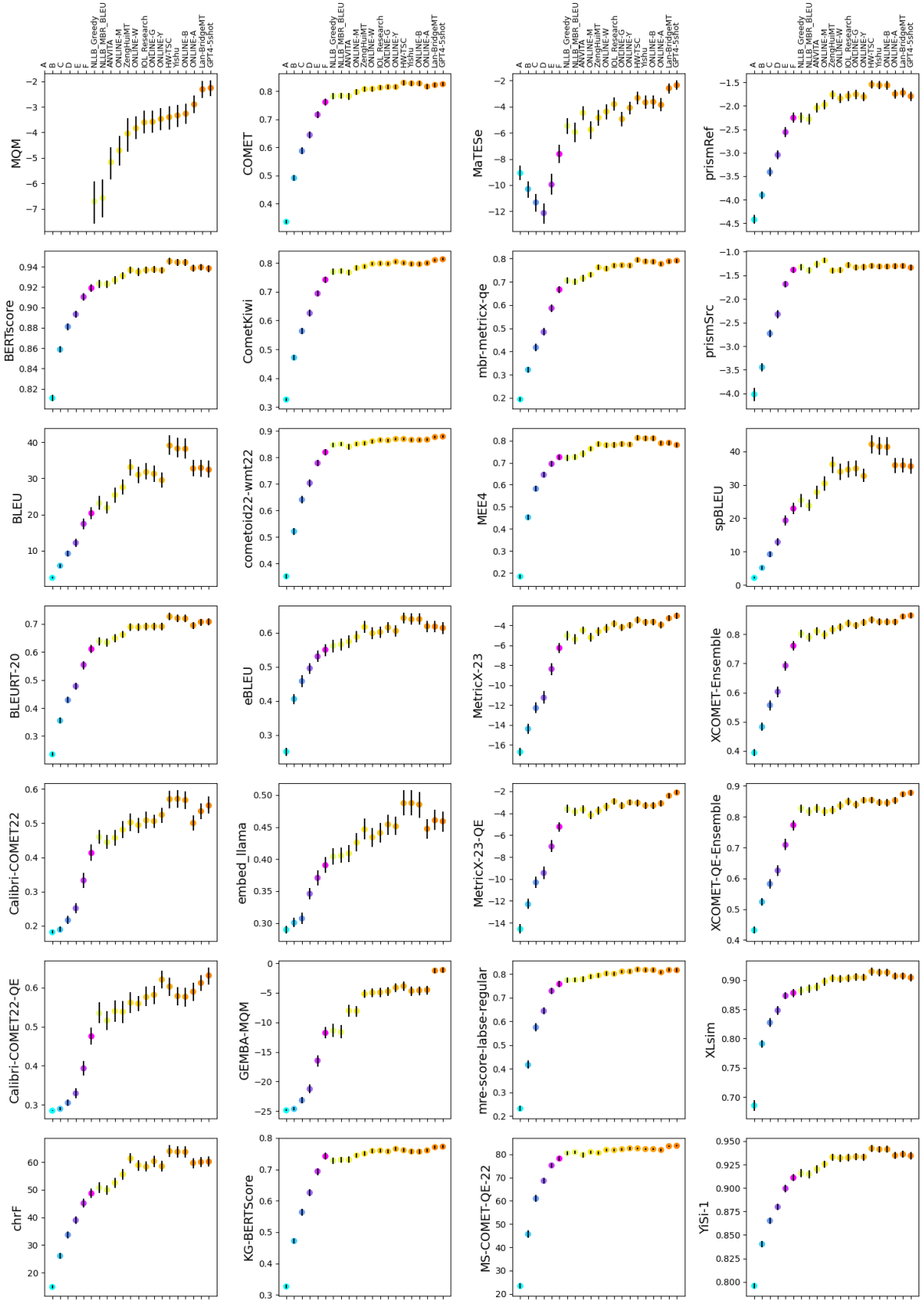

Figure 9: System average scores (with error bars computed via bootstrap resampling 1000 times for $p < 0.05$) for ZH→EN across the challenge set (cool colours/left) and submitted WMT systems (warm colours/right). Our challenge set systems are ordered from left to right with BLEU scores, while the submitted WMT systems are ordered by MQM score on the news domain.

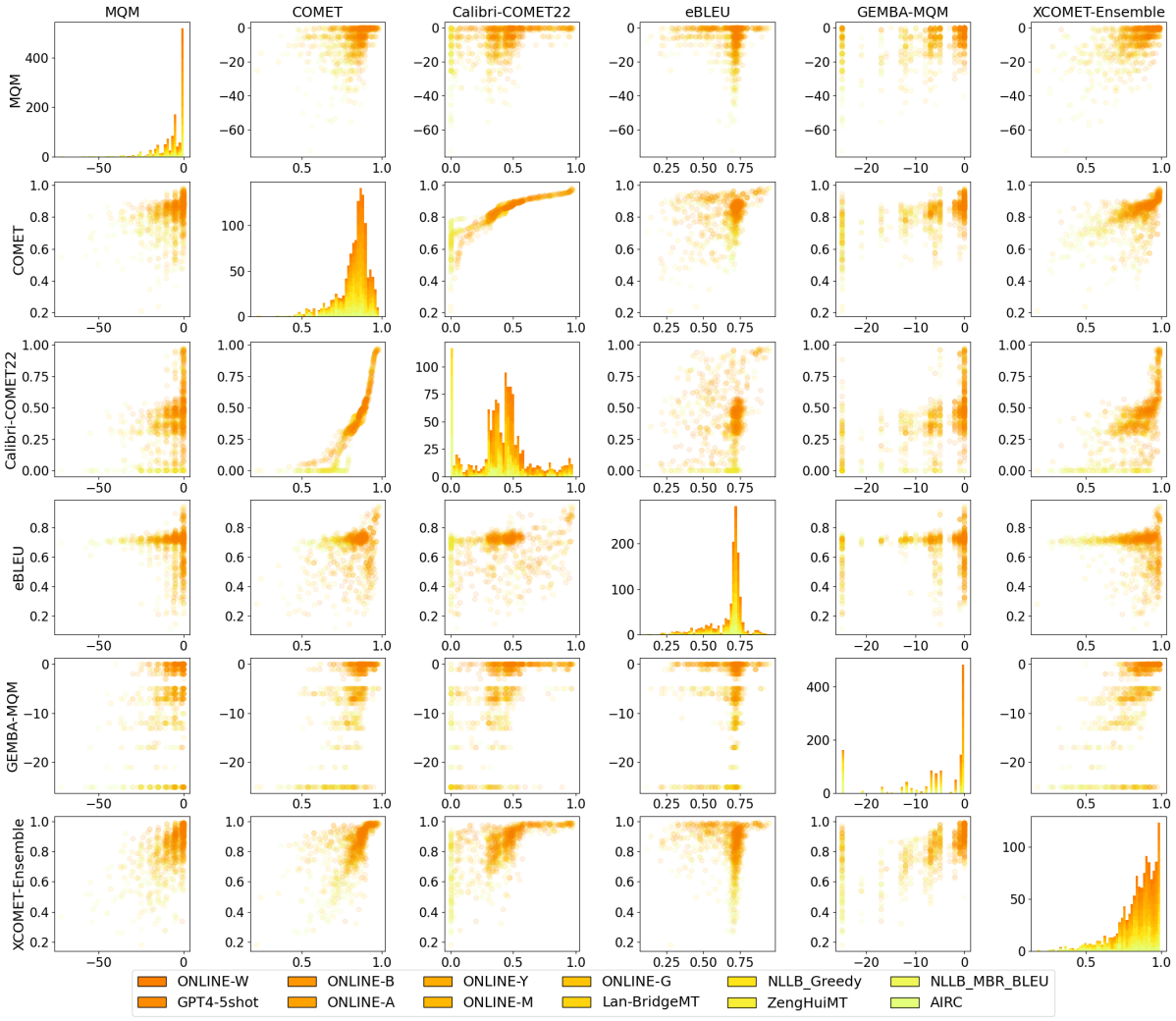

Figure 10: A subset of the metrics (and MQM scores) for EN→DE, showing only the high-quality WMT MT system submissions. The diagonal entries show stacked histograms of segment scores. The off-diagonal entries are scatterplots where each point is a single segment positioned according to the score assigned to it by row and column metrics; each point is coloured according to the MT system that produced it.