Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies

Anonymous ACL submission

Abstract

Ten years ago a single metric, BLEU, governed progress in machine translation research. For better or worse, there is no such consensus today, and consequently it is difficult for researchers to develop and retain the kinds of heuristic intuitions about metric deltas that drove earlier research and deployment decisions. This paper investigates the "dynamic range" of a number of modern metrics in an effort to provide a collective understanding of the meaning of differences in scores both within and among metrics; in other words, we ask *what point difference x in metric y is required between two systems for humans to notice?* We conduct our evaluation on a new large dataset, ToShip23, using it to discover deltas at which metrics achieve system-level differences that are meaningful to humans, which we measure by pairwise system accuracy. We additionally show that this method of establishing delta-accuracy is more stable than the standard use of statistical p-values in regards to testset size. Where data size permits, we also explore the effect of metric deltas and accuracy across finer-grained features such as translation direction, domain, and system closeness.

1 Introduction

A decade ago, the BLEU metric served as the default metric for machine translation evaluation. It was not without its criticisms (Hovy and Ravichandran, 2003; Callison-Burch et al., 2006; Belz and Reiter, 2006) or compelling alternatives (Banerjee and Lavie, 2005; Popović, 2015), but a combination of adequate performance, robustness to new languages, simplicity and understandability, and inertia helped it retain this position. This is no longer the case. BLEU's deficiencies quickly became apparent as deep learning approaches to machine translation replaced the earlier symbolic paradigms (Mathur et al., 2020a). Today, a number of metrics—themselves deep-learning based—compete in an ecosystem where there is no longer any dominant, default metric.

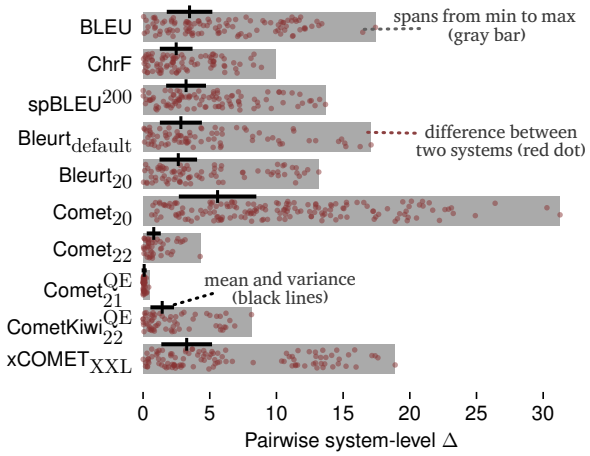

Figure 1: Distribution of pairwise system deltas for each metric over all systems from WMT22. Gray rectangles show min-max range which is vastly different between metrics. Standard deviations (black lines) also differ.

This situation creates a problem for researchers working to keep abreast of developments in the field. Different metrics, including different models within the same metric family, have different "dynamic ranges", i.e., the range of scores one can expect to see. Furthermore, the "metric delta", i.e., the score difference signifying a meaningful change in performance between two systems, also varies across metrics. It is perhaps understandable that some practitioners therefore continue to use BLEU, as well, if only to ground their understanding.

This paper attempts to introduce some order and clarity into this situation. We make use of a large, new human evaluation dataset, ToShip23, to compare the score ranges of metrics on a large number of systems against pairwise system-level accuracy. Importantly, we break down these accuracy scores into bins based on metric deltas, which allows us to determine accuracies for each metric as a func-

[0]Online tool: *Link retracted for anonymity*

tion of the score differences between two systems. This provides a measure of confidence in the output that is stable across testset size, in contrast to standard statistical significant testing, which becomes more stable as testset size grows. We release a tool that allows a user to easily compare accuracies at different threshold across metrics.[0]

2 Experimental Setup

Data. We perform experiments related to evaluation of MT outputs based on a proprietary dataset *ToShip23* which is of a magnitude larger than any publicly available data and enables more fine grained glimpse into the metrics behaviour. The dataset is an extended version of ToShip21 dataset (Kocmi et al., 2021) with details described in Appendix B. We also use data from the annual WMT evaluation campaigns to validate our results, specifically the metrics shared task (Freitag et al., 2022b, 2023), to make results replicable. We only use MQM (Freitag et al., 2021a) and DA+SQM (Kocmi et al., 2022) subset of human evaluated systems because reference-based DA (Bojar et al., 2016) is suboptimal for the evaluation of modern MT systems (Freitag et al., 2022b). See Table 1 for an overview of dataset sizes.

Dataset	Segments	Systems	Sys. pairs	Langs.	Domains
WMT22	221k	108	543	8	4
WMT23	223k	129	871	7	4
ToShip21	2300k	4380	3344	101	2
ToShip23	3016k	6752	6530	94	>10

Table 1: Sizes and coverage for the human annotated datasets used in this work.

Investigated Metrics. We evaluate the most frequently used metrics in machine translation: BLEU (Papineni et al., 2002), ChrF (Popović, 2015), sp-BLEU (Goyal et al., 2022), BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020).

BLEU and ChrF are n-gram matching heuristics while the rest uses a parametric model to produce a segment-level score of a translation. $\text{Comet}_{21}^{\text{QE}}$ and $\text{CometKiwi}_{22}^{\text{QE}}$ are special cases which do not require a reference. We do not include any LLM-based metrics (Fernandes et al., 2023; Kocmi and Federmann, 2023) which are not replicable because of non-publicly available models. Find the specific models, implementation details, and selection rationale in Appendix A.

Metric Delta. We focus solely on the pairwise system ranking – deciding which system is better based on a system-level score[1] difference between two systems. We refer to this as *metric delta* (Δ).

Because the order of two systems is arbitrary and do not affect any of our evaluations, only some visualizations, we always specify the first system to be the superior with respect to the human judgement. Therefore human score delta is always positive, while metric delta may be negative, which represents that metric incorrectly ranked the system pair in reverse order to humans.

Pairwise Accuracy. To test the correlations between automatic metrics and human judgement, we use pairwise accuracy (Kocmi et al., 2021): how many system pairs does the metric rank the same way as humans over the total number of system pairs in the dataset. Formally:

$$\text{Acc} = \frac{|\text{sign}(\text{metric}\Delta) = \text{sign}(\text{human}\Delta)|}{|\text{all system pairs}|} \quad .$$

3 Unifying Metric Ranges

We first look at the "dynamic ranges" exhibited by different metrics across our datasets. We ground these deltas in human scores by comparing pairwise system-level accuracy at different thresholds of delta. With this, we are able to establish a table of average metric deltas for different accuracy levels, and build a simple model that maps any metric into the unified space of estimated accuracies.

3.1 Various Ranges for Metric Deltas

Figure 1 depicts the distribution of system-level score deltas for various metrics. Some metrics have similar ranges, such as ChrF and BLEU, while others use much larger score range (Comet_{20} has $\sim 5\times$ higher deltas to BLEU) or lower score range ($\text{Comet}_{21}^{\text{QE}}$ has $\sim 1/5$ range of BLEU).

It may be tempting to attempt to bring together these score ranges onto a single scale, e.g. by linear interpolation, perhaps towards BLEU scale. We begin by noting the infeasibility of arriving at a single unified scale. We use the mostly unwritten (and long-debunked (Mathur et al., 2020a)) operating assumption that +1–2 BLEU points denotes a significant finding as an anchor point to illustrate the range of metric deltas on a subset of systems in Figure 2. This figure reports metric deltas for six randomly-selected system pairs from WMT23 data, whose delta was roughly 1 BLEU.

[1] Usually the average of all segment level scores.

	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 6
BLEU	1.0	1.0	1.0	-1.0	-1.0	-1.0
ChrF	1.2	0.5	3.1	-0.4	-0.3	5.9
spBLEU^{200}	1.2	2.1	5.0	-0.6	-0.9	5.3
$\text{Bleurt}_{\text{default}}$	2.4	0.1	-0.5	-0.5	-0.2	8.6
Bleurt_{20}	1.2	2.3	2.9	1.6	-0.6	8.5
Comet_{20}	1.3	11.1	2.3	6.8	-3.4	16.3
Comet_{22}	0.1	2.1	0.7	0.6	-0.6	1.9
$\text{Comet}_{21}^{\text{QE}}$	0.0	0.2	0.0	0.1	-0.1	0.2
$\text{CometKiwi}_{22}^{\text{QE}}$	0.9	3.3	0.4	1.5	-0.4	4.3
$\text{xCOMET}_{\text{XXL}}$	2.4	3.7	1.7	4.3	-1.2	10.0
Human	Accept	Accept	Accept	Accept	Accept	Accept

Figure 2: Subset of system pairs from WMT23 that have ∼1 BLEU delta. Each column is one system-pair. Dark background represent metric disagreeing with humans on system ranking. This highlights that normalizing metrics towards BLEU range is not feasible.

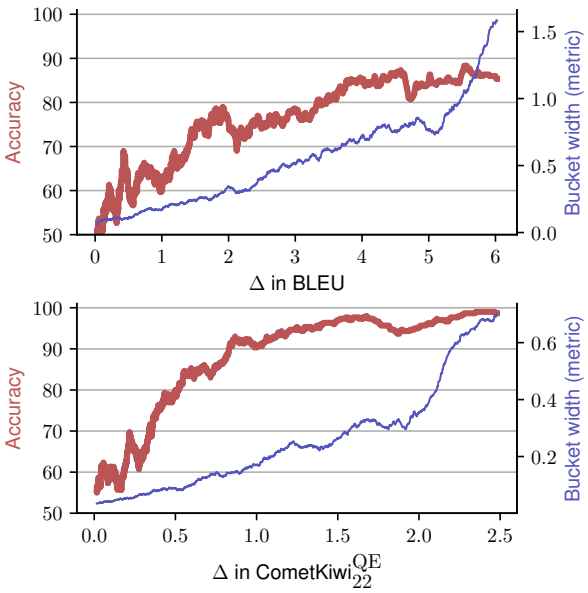

Figure 3: What pairwise accuracy (left-y-axis) to expect when seeing given certain acceptance threshold (x-axis). The bin width (right-y-axis) shows the width of the bin for metric delta that contains 300 system pairs.

In addition to the wide ranges of scores, we also observe that metrics do not always have the same direction or agreement with human judgment. Reconciling these by projection is not possible, due to an obvious point: metrics differ not just in the range of their scores, but in their accuracies. To better understand the problem, we look next into what are the implications of different thresholds. Specifically, we investigate how different delta correspond to humans being able to differentiate systems.

3.2 Accuracy of Metric Deltas

Many factors affect metric behavior:

- Each metric and also humans weights various phenomena differently, especially fluency versus adequacy (Amrhein et al., 2022).
- The reliability of Metrics differs when compared to humans (Mathur et al., 2020b; Freitag et al., 2021b, 2022b, 2023; Kocmi et al., 2021).
- Reference-based metrics are affected by quality of human references (Freitag et al., 2023; Zouhar and Bojar, 2024).

The pairwise accuracy as usually reported (Kocmi et al., 2021; Freitag et al., 2023) represent a value over the full dataset for all system pairs metric deltas. It does not take into consideration the *size* of the delta between systems, which heavily affects the accuracy; that is, whether the metric gap between two systems was large or small. However, this information is important in establishing equivalency of deltas across metrics.

To investigate this, we use a binning approach on the ToShip23 testset. Pairwise system deltas are sorted, and for each delta level, we group the closest 300 pairs into a same bin. For each bin, we plot the mean delta for that bin against the system-level pairwise accuracy.[2]

Figure 3 depicts this information for both BLEU and $\text{CometKiwi}_{22}^{\text{QE}}$. The red line shows that we need around 1.3 BLEU delta to reach 70% pairwise accuracy and 3.5 BLEU to reach 80% accuracy against the human judgments. Because BLEU is not a reliable metric, it never reaches 90% accuracy with humans, even for deltas as high as 6 BLEU points. In contrast, $\text{CometKiwi}_{22}^{\text{QE}}$ reaches 90% accuracy already at around 0.9 points and gets close to 100% accuracy past 2 $\text{CometKiwi}_{22}^{\text{QE}}$ points.

Our use of fixed-size bins introduces a caveat into the evaluation. Because our data points do not have a uniform delta distribution, the "width" of each bin (defined as the difference between the smallest and largest delta) grows as we move towards larger deltas, where data points are sparser. This width is depicted by the blue line in Figure 3. As we increase the delta, there are fewer and fewer systems with as large delta and thus we need to take system pairs that are farther from the investigated delta. For example, for calculating the pairwise accuracy of 1 BLEU point, we take system pairs with a delta of 1 ±0.1 (half of 0.2), while for 3 BLEU the width of a bin is 3 ±0.25 points. The

[2] Appendix C investigates other sizes of bins than selected 300 system pairs.

Estimated Accuracy	Coin toss 50%	55%	60%	65%	70%	75%	80%	85%	90%	95%
BLEU	0.27	0.52	0.78	1.06	1.39	1.79	2.34	3.35	-	-
ChrF	0.14	0.33	0.54	0.76	1.00	1.28	1.63	2.12	3.05	-
spBLEU^{200}	0.25	0.52	0.82	1.13	1.49	1.91	2.46	3.28	5.57	-
$\text{Bleurt}_{\text{default}}$	0.23	0.66	1.11	1.59	2.11	2.71	3.43	4.39	5.98	-
Bleurt_{20}	0.02	0.17	0.33	0.49	0.66	0.85	1.07	1.35	1.73	2.44
Comet_{20}	0.08	0.36	0.65	0.96	1.29	1.67	2.10	2.66	3.45	5.10
Comet_{22}	0.03	0.10	0.18	0.26	0.35	0.45	0.56	0.71	0.94	1.53
$\text{Comet}_{21}^{\text{QE}}$	0.003	0.008	0.013	0.019	0.025	0.032	0.041	0.052	0.073	-
$\text{CometKiwi}_{22}^{\text{QE}}$	0.01	0.08	0.16	0.24	0.33	0.42	0.53	0.67	0.85	1.18
$\text{xCOMET}_{\text{XXL}}$	0.02	0.19	0.37	0.56	0.76	0.98	1.24	1.55	1.99	2.74

Table 2: Thresholds and estimated accuracies for each metric on ToShip23 dataset averaged across all language pairs. For example, when requiring 90% of decisions be the same as humans, improvement needs to be ≥3.05 ChrF, ≥0.85 $\text{CometKiwi}_{22}^{\text{QE}}$, and BLEU never reaches this accuracy threshold.

bin width mainly affects the tail of the evaluation.

As our evaluation is empirical, it is heavily affected by the underlying systems and the lines fluctuate. In the next section, we try to fit a smooth line to abstract the results, followed by discussion which phenomena affect the pairwise accuracy.

3.3 Aligning Metrics on Accuracy

Practitioners might be interested in getting an intuition behind a particular metric delta, e.g. +0.10 of Comet_{22} and how such delta relates to other metrics that they are familiar with. Clearly, the higher the delta, the more likely that human raters would also notice the quality difference between systems. It remains unclear what delta is enough to warrant acceptance. To this end, we use the estimated accuracy results introduced in previous subsection. As the estimated accuracy line is noisy, we fit a curve through the data and use it to derive thresholds for comparing various metric deltas.

We use a parametrized sigmoid to fit a curve through the data. The choice of the sigmoid function is arbitrary and based on visual similarity and the feature that it converges towards fixed point and thus is bounded. This is a desired feature representing that each metric has a different overall reliability. We parameterize it using two variables φ and fit it with damped least square algorithm (Levenberg, 1944). The function is defined as:

$$f(x) = \frac{\varphi_1}{1 + \exp(-\varphi_2 \cdot x)} \quad .$$

The resulting fit is visualized in Figure 4. Although not perfect, it offers insight into the metric delta behaviour, specifically comparing different different deltas' estimated accuracy. We use the sigmoid functions to calculate estimated accuracy

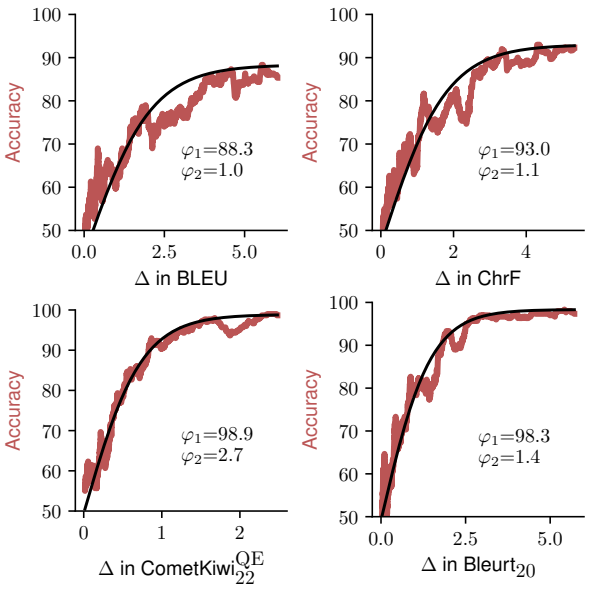

Figure 4: Empirical pairwise accuracies for various metrics with a fitted sigmoid curves on ToShip23 dataset. All metrics are in Figure 11.

for various levels of delta in Table 2. This is the core result of our work and helps in understanding how different metrics compare to each other.

For example, an improvement of 1.06 BLEU has the same estimated accuracy (65%) as the 0.24 $\text{CometKiwi}_{22}^{\text{QE}}$, while 3.35 BLEU has the same estimated accuracy as 0.67 $\text{CometKiwi}_{22}^{\text{QE}}$. And +1 improvement on $\text{CometKiwi}_{22}^{\text{QE}}$ signalize that in >90% scenarios, human annotators would agree with the ranking of $\text{CometKiwi}_{22}^{\text{QE}}$, while BLEU never reaches this level of agreement. Note that estimated accuracies are empirical from a given ToShip23 dataset. Therefore, we do not claim that +0.56 Comet_{22} yields 80% accuracy for all scenarios but rather that it is as accurate as +2.34 BLEU.

As these thresholds are combined for all scenar-

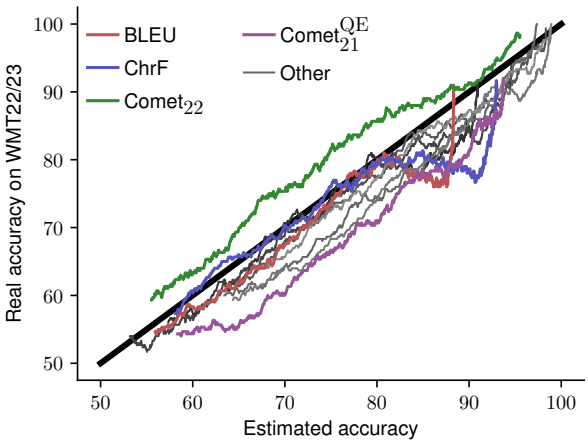

Figure 5: Testing the validity of thresholds devised on ToShip23 with WMT datasets. In a scenario without noisy data, we would expect the real accuracies to match the estimated accuracies (the black line). See detailed per-metric breakdown in Figure 10.

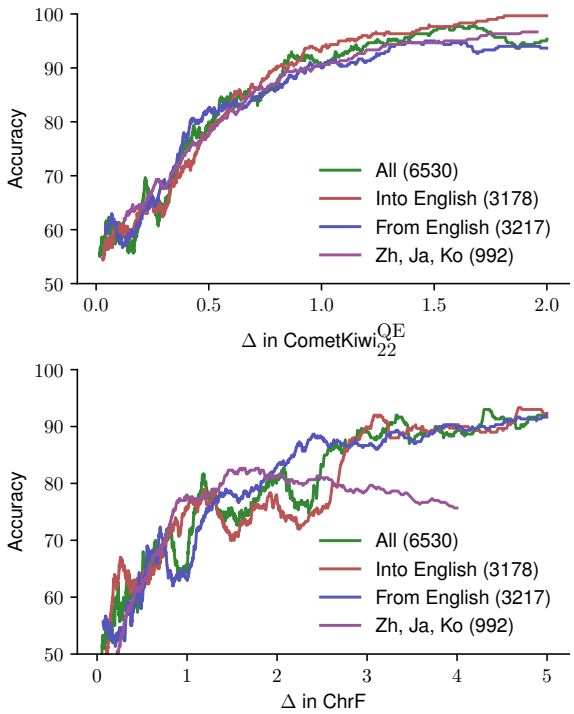

Figure 6: The comparison of pairwise accuracy on ToShip23 dataset when comparing into English, out-of-English, and Chinese, Japanese, Korean language pairs separately. The count shows total number of system-pairs in the evaluation. See other metrics in Figure 12.

ios, we dive in the next section into validating out results on public WMT dataset, followed with investigation of what affects the metric delta and how reliable the comparison is in different settings.

4 Factors Affecting Metric Deltas

We have *empirically* derived the estimated accuracy for various metrics. In this section, we investigate factors that affect metric delta and show how reliable the thresholds remain under these factors. These include the testset size, dataset and domain selection, and translation direction.

Additional factors could influence the metric delta, but we lack the data to evaluate these aspects. A key consideration is whether the metric delta is contingent on the underlying absolute values. In other words, we need to determine if a +1 BLEU delta varies in reliability based on these absolute values. For instance, does the impact of moving from 20 to 21 BLEU differ significantly from a shift from 60 to 61 BLEU in different system pairs?

4.1 Different Domains and Datasets

We derived the thresholds from ToShip23. Now, we validate them on WMT data to show how well they transfer. To address the relatively small size of WMT, we first combine the WMT 2022 and 2023 datasets yielding 1414 system pairs. This dataset contains different set of segment sources and domains, and was evaluated with mix of MQM and DA+SQM human evaluation protocols. In order to test the thresholds, we take scores for all WMT system pairs and convert them into estimated accuracies via devised thresholds. For each estimated accuracy level, we take the closest 300 system pairs and calculate the real accuracy on WMT data. If the mapping would be perfect and we had enough samples, the estimated accuracy would match the real accuracy for each investigated level.

We show the results in Figure 5. In the ideal case, we would expect the real accuracies and estimated accuracies to match, however, the noise from empirical data affects the results. Some metrics are consistently underestimated, such as Comet$_{22}$ which has higher real accuracies on WMT dataset that the estimated accuracies. On the other hand, Comet$_{21}^{QE}$ has much lower accuracies on WMT data and our thresholds overestimate it.

Overall, the trend is clear and the thresholds normalize all metrics into a shared space of estimated accuracies. Therefore, we advise reporting accuracy when presenting results, together with significance testing and metric delta.

4.2 Language Pair

Notoriously, a large gap in absolute BLEU scores exists between languages (Denoual and Lepage, 2005; Post, 2018). This reflects properties like data

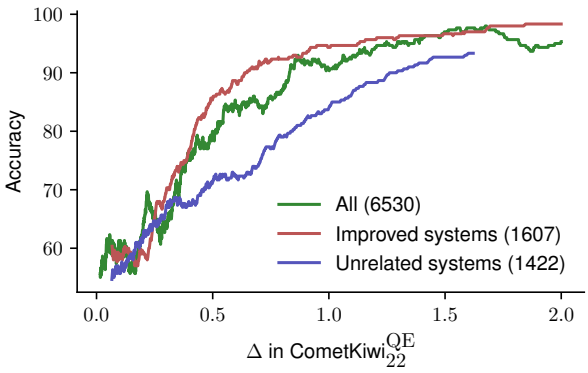

Figure 7: The comparison between iterated and unrelated systems on ToShip23. See other metrics in Figure 13.

sizes, attention progress in different languages, and target-side morphological complexity.

Unfortunately, there is not enough data to examine each language pair individually. Instead, we bin languages into two groups, *into-English* (XE) and *out-of-English* (EX) language pairs, which does leave us with enough data in the ToShip23 dataset. In addition, we separate system pairs containing Chinese, Japanese, or Korean (CJK) together.

Figure 6 show the accuracy with a subset of system pairs depending on a languages. There is some fluctuation between XE and EX, but the behaviour is comparable. This is interesting, since most of the underlying testsets have authentic source (e.g., not using testset in reverse direction, Toral et al., 2018). The CJK group does also perform similarly for CometKiwi$_{22}^{QE}$, but not for ChrF. This shows the thresholds are invalid for all metrics and scenarios and are affected by whether metrics evaluate all language similarly or not.

4.3 Iterated versus Unrelated Systems

Another main difference that affects the evaluation is if the systems are closely related. One point of distinction is between *iterated systems* (a baseline against specific improvements, produced by the same group) or *unrelated system* (for example, WMT yearly evaluation which comes from different teams). It has long been known that surface metrics like BLEU work best when evaluating closely related systems (Callison-Burch et al., 2006). It may be easier for both metrics and humans to distinguish improved system over baseline, while comparing unrelated systems adds a difficulty of weighting different styles and errors.

To investigate this, we use the labels of ToShip23 dataset, where some system pairs are improvements over the previous model, while other system pairs are completely unrelated and developed by different teams, similarly to WMT evaluation. Figure 7 confirms the assumption, that unrelated systems are harder to evaluate and that the metric behaves differently. Therefore, automatic metrics are better to rank improved systems which is mostly their use case than unrelated systems. Importantly, some metrics such as BLEU have much harder time to distinguish unrelated systems, see Figure 13 in Appendix. This effect should be investigated to larger detail in the future works.

4.4 Testset Size

Another phenomena that may affect the system delta is the number of sentences in the parallel testset used to evaluate pair of systems. Common wisdom says that the testset should be as large as possible. We ask if increasing testset size affects the system delta and its statistical significance.

To examine how testset size affects the metric delta, we take a system pair and sample testsets with increasing number of sentences. For each sample, we calculate CometKiwi$_{22}^{QE}$ delta and p-value using paired Student's t-test (Mathur et al., 2020a). We sample with repetition various testset sizes. For each testset size, we plot the average metric delta (or p-value respectively) over 50 runs together with the confidence interval.

From Figure 8, the metric delta fluctuates but keeps being mostly constant. The variance of the metric delta is higher for small testset sizes (under 500 segments). On the other hand, the p-value associated to the comparison hypothesis goes down simply by having a larger testset. The p-value goes down faster for larger effect sizes.

This is a natural phenomenon of statistical significance testing (Greenland et al., 2016). P-values decrease with an increasing sample size, assuming the null hypothesis does not hold. This is due to the increase in statistical power—the probability that the test correctly rejects the null hypothesis when it is false. Should the null hypothesis hold perfectly, which is rarely the case, increasing the sample size would not systematically affect the p-values. Therefore, it is possible to claim a statistically significant improvement over a baseline model even with a small metric delta, which might not be noticeable by humans, just by using a large-enough testset. This conclusion is not an argument against the use of statistical significance testing, which is important especially when observing smaller deltas.

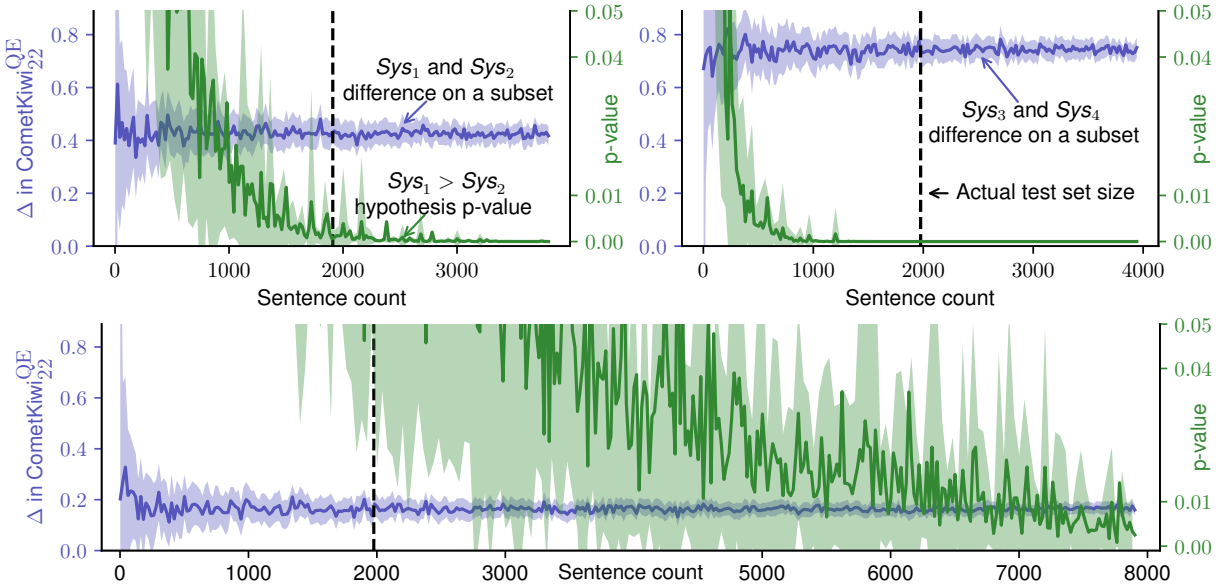

Figure 8: Three system pairs on different languages from WMT23 scored by CometKiwi$_{22}^{QE}$. The blue line is the average system delta for given testset size and green line is the associated p-value. Values to the right of the dashed line are supersampled and shaded areas are 99.9% confidence intervals from t-distribution. The metric delta does not change much while the p-value goes down with higher subset size.

An interesting research question is if we could fix the testset size to stabilize p-value.

Overall, this shows that metric delta is stable under different testset sizes, while statistical significance testing is affected by it. We assumed to be adding sentences from the same distribution. The metric delta can be manipulated by adding segments that are more difficult than the rest.

5 Discussion

5.1 Best-performing Metrics

With the ToShip23 dataset, we can also calculate total pairwise accuracy over all system pairs to devise which metric perform the best on up to date the largest dataset of MT human evaluation. We follow the same evaluation as in Table 2 from Kocmi et al. (2021). With twice as large dataset than ToShip21, extended by state-of-the-art systems from 2022 and 2023, we can see how metric perform on system-level ranking. Table 3 shows that the best performing metric over the ToShip23 dataset is CometKiwi$_{22}^{QE}$ by a small margin over xCOMET$_{XXL}$. CometKiwi$_{22}^{QE}$ is a quality estimation metric, which has an additional bonus of not being affected by reference bias.

Additionally, we can notice the overall accuracy dropped for all metrics in the last two years. This may have several possible explanations and likely does not mean drop in metrics performance:

	ToShip23	**22-23**	**19-21**	**WMT23**
system pairs (N)	6530	1843	4687	249
CometKiwi$_{22}^{QE}$	81.5	74.5	84.3	90.0
xCOMET$_{XXL}$	81.4	75.3	83.9	92.8
Comet$_{20}$	80.1	73.2	82.9	86.3
Bleurt$_{20}$	78.6	69.8	82.1	89.2
Comet$_{22}$	78.6	71.1	81.5	84.7
Comet$_{21}^{QE}$	76.8	71.2	79.0	69.5
ChrF	71.9	61.4	76.0	79.5
spBLEU^{200}	71.6	61.0	75.7	81.9
BLEU	70.3	61.3	73.9	81.5
Bleurt$_{default}$	69.9	61.0	73.4	85.1

Table 3: A pairwise accuracy over all system pairs from ToShip23 and two subsets depending on the year of evaluation. The results of MQM subset of WMT23 (Freitag et al., 2023).

- **Different systems**: Newer architectures or systems that are closer to each other in performance, thus harder to evaluate by humans
- **New testsets**: While the 2019-2021 contains only two domains, the newer data have been evaluated on a much larger set of domains, where some domains may be challenging for metrics
- **Human bias**: The protocol changed over years which may shifted annotator's scoring patterns.

However, the absolute pairwise accuracy is less important than the ranking of metrics, as it is heavily affected by the system pairs. We compare to MQM subset of Freitag et al. (2023), which ranks metrics

in similar order supporting our findings. There are some notable differences, such as $\text{Comet}_{21}^{\text{QE}}$ ranking as the worst metric in WMT, while $\text{BLEURT}_{\text{default}}$ is the worst in ToShip23. Since many aspects of the evaluation are different, we do not dive into comparison, but rather highlight the overall picture. ToShip23 corroborates that QE metrics have reached the quality of reference-based metrics, as well as the (already well-established) fact that lexical-based metrics are not useful for evaluating high-resource MT models these days.

5.2 Recommendations for MT Evaluation

We conclude with a list of recommendations for automatic MT evaluation:

- Use $\text{CometKiwi}_{22}^{\text{QE}}$ as the main metric. In addition to its better performance, as a quality estimation metric, it is not affected by references.
- Use at least one additional metric of a different type; e. g. BLEURT_{20}, which is reference-based and uses a different architecture from Comet.
- For each metric delta, report estimated accuracy to help align reliability of used metrics.

In addition, caution is warranted when employing the same metric for evaluation that was utilized during training, as this practice may lead to artificially inflated results. For instance, it is advisable not to evaluate with the same metric used for Minimum Bayes Risk Decoding (Freitag et al., 2022a), QE metric used for corpus filtering (**?**), or avoid using metrics built on the same model as the translation system because LLMs tend to favor outputs generated by themselves (Liu et al., 2023).

6 Related Work

The closest work to ours is Lo et al. (2023), who investigate the relationship between metric deltas and the p-value of human ranking, concluding that not even 2 BLEU points reliably correspondent to human judgement. This aligns with our work that two BLEU points reaches an estimated accuracy of only 77.2%. Their work also does not consider the directionality of the delta, and consequently they do not penalize situations where humans and metric disagree on which system is better.

Mathur et al. (2020a) found that even statistical significant deltas of up to three BLEU points do not reliably correspond to human judgement. In a broad survey, Marie et al. (2021) notes that various community "rules of thumb" about sufficient BLEU deltas might be the result of an evolved consensus that has no basis in scientific evidence. Similarly, Kocmi et al. (2021) demonstrated that among system pairs deemed statistically significant by humans and where BLEU disagree with humans, the median delta is 1.3 BLEU. Marie (2022) reinvestigated the WMT 2020 and 2021 results and showed that deltas lower than 2 BLEU needs to be tested for statistical significance.

Automated metrics in NLP and MT have been under scrutiny for long time. Hovy and Ravichandran (2003) raised early doubts about BLEU. Callison-Burch et al. (2006) pointed to failure modes of BLEU and suggested it be used in more narrow situations. Post (2018) identified a problem with conflicting implementations of BLEU and offered a unified solution. The broader field of computer science has been concerned with what is a meaningful acceptance threshold of a metric (Mori et al., 2018). The acceptance thresholds are usually established to trade off risks in types of errors (Shatnawi et al., 2010). Kelley and Preacher (2012), studying effect sizes in psychology, summarize that effect sizes should be scaled appropriately. Alike, Plonsky and Oswald (2014) ask what effect size suffices and note its dependence on the variance and that all acceptance thresholds are arbitrary.

7 Conclusion

In this work, we investigated the interpretation of deltas from automatic machine translation metrics. Although metrics have different ranges of scores, what ultimately matters to the practitioner is how score *deltas* are grounded in human ability to perceive those differences, which we judge by pairwise system-level accuracy on a large collection of human judgments. We empirically determined thresholds for popular metrics to align them on accuracy and provide a tool[3] that relates metrics to each other. Finally, we showed the importance of using metric-delta accuracy over p-values: the former is stable across testset sizes.

We undertook some investigations into subfactors of the data, showing that the results were robust to, for example, translation direction, and also that they generalized to different testsets. These investigations were limited by the data size. For future work, it would be useful to explore delta-accuracy for different subsets and combinations of features, presuming that enough data were available for the task.

[3] *Link retracted for anonymity.*

Limitations

While this work provides more informed guidelines on interpreting metric delta, they remain crude and do not fix the inadequacy of automated metrics. In order to guarantee improvements, human evaluations need to be carried out.

We use humans as a gold standard, however, they are noisy and also unreliable especially for systems that are close in performance.

Our estimated accuracy should not be used as the reason to reject a result, similarly as low significance p-value.

Ethics Statement

The human annotators have been compensated considerably higher than the minimum wage standards in their respective countries. This commitment reflects our dedication to fair labor practices and the well-being of those contributing to our work.

References

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.

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.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Etienne Denoual and Yves Lepage. 2005. BLEU in characters: Towards automatic MT evaluation in languages without word delimiters. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022b. 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. 2021b. 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.

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

Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. 2016. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31:337–350.

Eduard Hovy and Deepak Ravichandran. 2003. Holy and unholy grails. In *Proceedings of Machine Translation Summit IX: Plenaries*, New Orleans, USA.

Ken Kelley and Kristopher J Preacher. 2012. On effect size. *Psychological methods*, 17(2):137.

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. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. 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.

Kenneth Levenberg. 1944. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2023. Llms as narcissistic evaluators: When ego inflates evaluation scores. *arXiv preprint arXiv:2311.09766*.

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.

Benjamin Marie. 2022. Yes, we need statistical significance testing.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. 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.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Allan Mori, Gustavo Vale, Markos Viggiato, Johnatan Oliveira, Eduardo Figueiredo, Elder Cirilo, Pooyan Jamshidi, and Christian Kastner. 2018. Evaluating domain-specific metric thresholds: an empirical study. In *Proceedings of the 2018 International Conference on Technical Debt*, pages 41–50.

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.

Luke Plonsky and Frederick L Oswald. 2014. How big is "big"? interpreting effect sizes in l2 research. *Language Learning*.

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.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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.

Raed Shatnawi, Wei Li, James Swain, and Tim Newman. 2010. Finding software metrics threshold values using roc curves. *J. Softw. Maint. Evol.*, 22(1):1–16.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Vilém Zouhar and Ondřej Bojar. 2024. Quality and quantity of machine translation references for automated metrics.

A Metric Implementation Details

There are many automatic metrics that has been developed. In our study, we focus only on the most used or the best performing ones. Here is the description, reason for their selection and details of implementations used. For metric quality, please, refer to Section 5.1 or Kocmi et al. (2021); Freitag et al. (2023).

Out of the lexical-based metrics, we select three of them, which are the most used. However, we emphasize that these metrics should no longer be used for MT evaluation (Freitag et al., 2022b). We use SacreBLEU (Post, 2018) in version 2.3.1 with default setting:

- BLEU (Papineni et al., 2002): the most popular and currently one of the worst performing MT metrics (we used a specific tokenizer for Japanese and Chinese as recommended)
- ChrF (Popović, 2015): second most popular lexical-based metric with better performance
- spBLEU^{200} (Goyal et al., 2022): metric popular when evaluating on Flores testset

Two BLEURT models (commit `cebe7e6`):

- BLEURT$_{\text{default}}$ (Sellam et al., 2020): the default model when using BLEURT framework called BLEURT-Tiny. It is important to note, that its performance is worse than BLEU (Section 5.1) and should not be used as authors suggest.
- BLEURT$_{20}$ (Pu et al., 2021): the best performing Bleurt model

We evaluate five Comet models (`v2.1.0`), the most popular metric framework aside BLEU:

- Comet$_{20}$: most frequently used model and the default reference based model until the end of year 2023. The model name `wmt20-comet-da`.
- Comet$_{22}$: currently the default reference-based model (`wmt22-comet-da`), outperforming Comet$_{20}$.
- Comet$_{21}^{\text{QE}}$: we picked `wmt21-comet-qe-mqm` for its unusual behaviour of using very small delta while reaching high pairwise accuracy.
- CometKiwi$_{22}^{\text{QE}}$: `wmt22-cometkiwi-da` is the best quality estimation model.
- xCOMET$_{\text{XXL}}$: the best performing publicly available metric as evaluated by Freitag et al. (2023).

B ToShip23 Dataset Details

For this work, we introduce and utilize an extended version of a non-public ToShip23 dataset. The main changes of a dataset are almost twice as many system pairs as in ToShip21 (Kocmi et al., 2021); more than ten new domains and new parallel testsets; improved human evaluation protocol; and evaluating the latest state-of-the-art MT models.

The parallel testsets for evaluating MT models that we use in the extended part are mostly a collection of non-published human translated sentences. We focus on using testsets in authentic direction, from original source into human translated reference (avoiding reverse testsets whenever possible, Toral et al., 2018). In contrast to ToShip21, which uses mainly two domains (news and speech), we extended the domains by more than ten.

We reduced the total number of languages in the ToShip23 from 101 to 94. The removed languages

are those which are not supported by either BERT (Devlin et al., 2019) or XLM-RoBERTa (Conneau et al., 2020) – language models used in the most popular metrics – therefore, we could not include those languages in our analysis.

The MT systems being part of the dataset are coming from the same distribution as in ToShip21, but evaluating the most recent state-of-the-art models including a limited number of LLM based translations. Lastly, we improved the human evaluation protocol, moved from source-based DA towards DA+SQM (Kocmi et al., 2022).

C Number of System Pairs in a Bin

In our work, we fixed the number of systems in a bin for given metric delta to 300 system pairs. We now show how this decision affected our evaluation. To this end, we show various bin sizes in Figure 9. The bin width works as a smoothing parameter. With bin size of 100 system pairs, the curve fluctuates, especially as one system pair transfer into 1% change on the accuracy scale.

We set the parameter to 300 system pairs because that is already a smoother curve, while not too wide so that the epsilon around the investigated delta is also not too high. However, this parameter should be re-investigated in the future works.

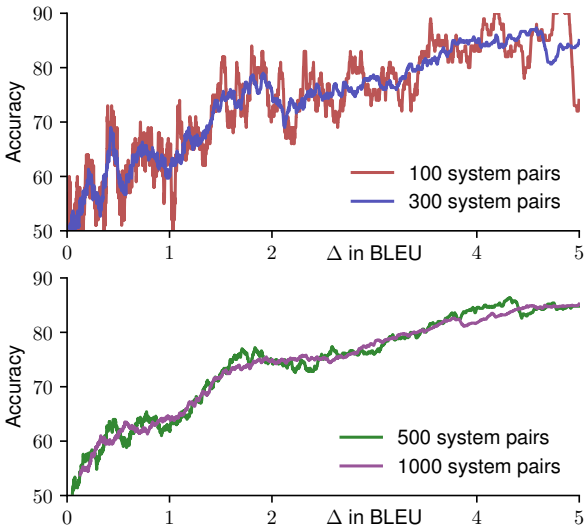

Figure 9: The comparison of pairwise accuracy for BLEU on ToShip23 dataset when we change how many system pairs are in evaluation for each individual delta.

D Extending Plots to All Metrics

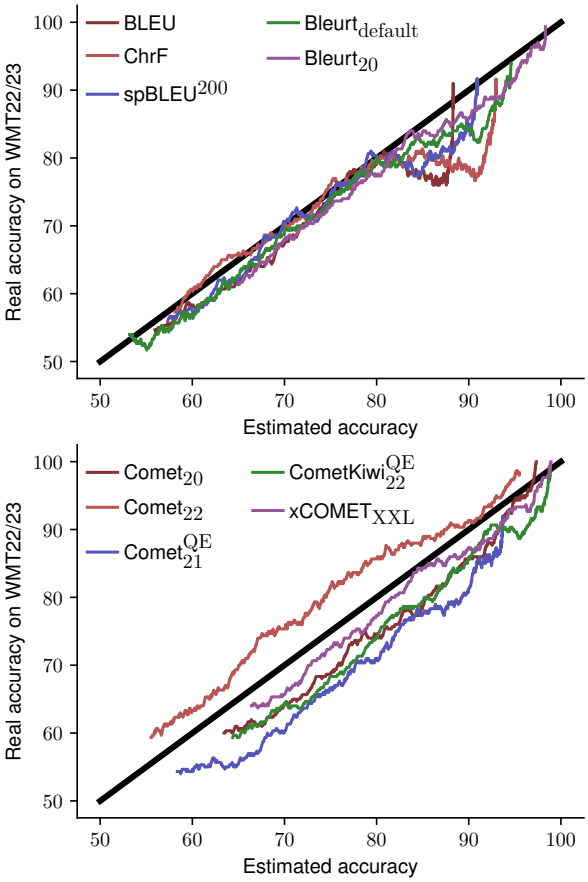

Figure 10: Testing the validity of thresholds devised on ToShip23 with WMT datasets. In a scenario without noisy data, we would expect the real accuracies to match the estimated accuracies (the black line). This figure provides more detail on Figure 5.

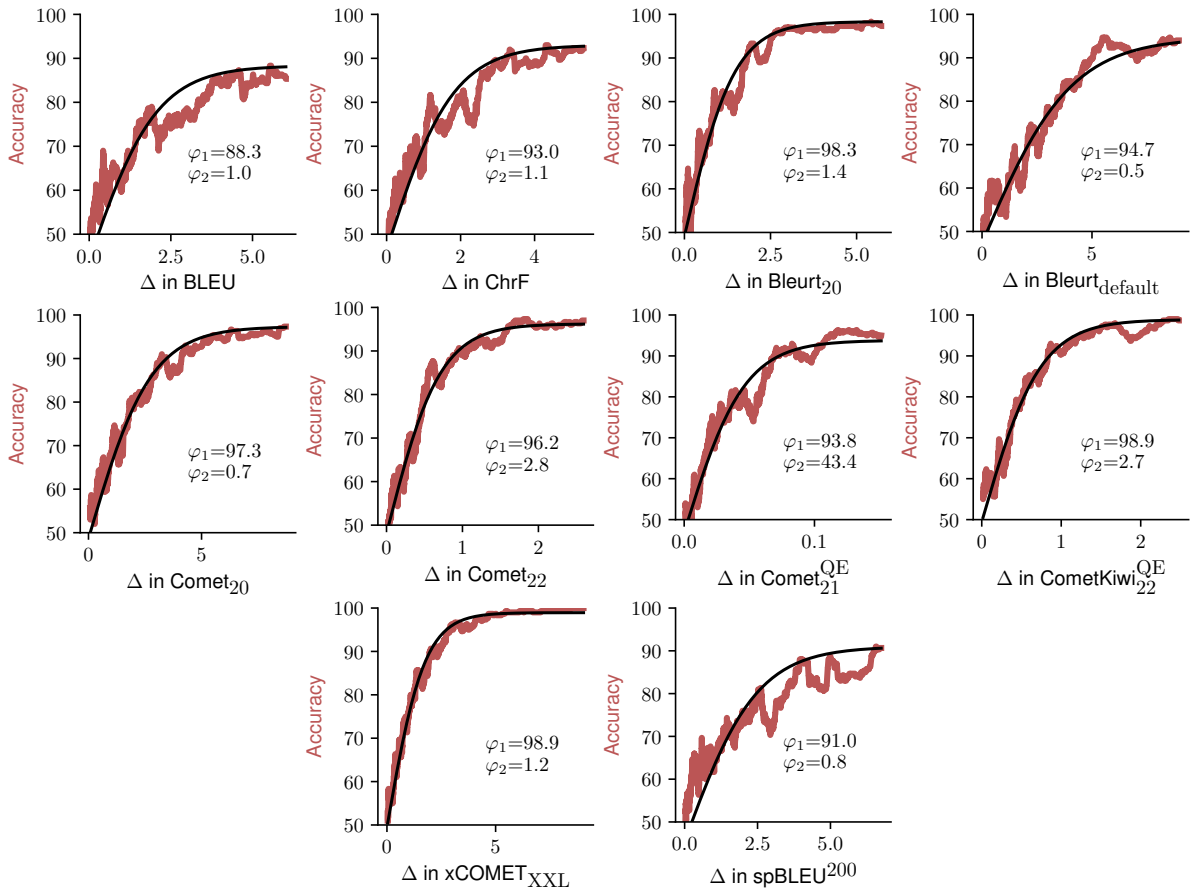

Figure 11: Empirical pairwise accuracies for all metrics with a fitted sigmoid curves on ToShip23 dataset. This figure extends Figure 4.

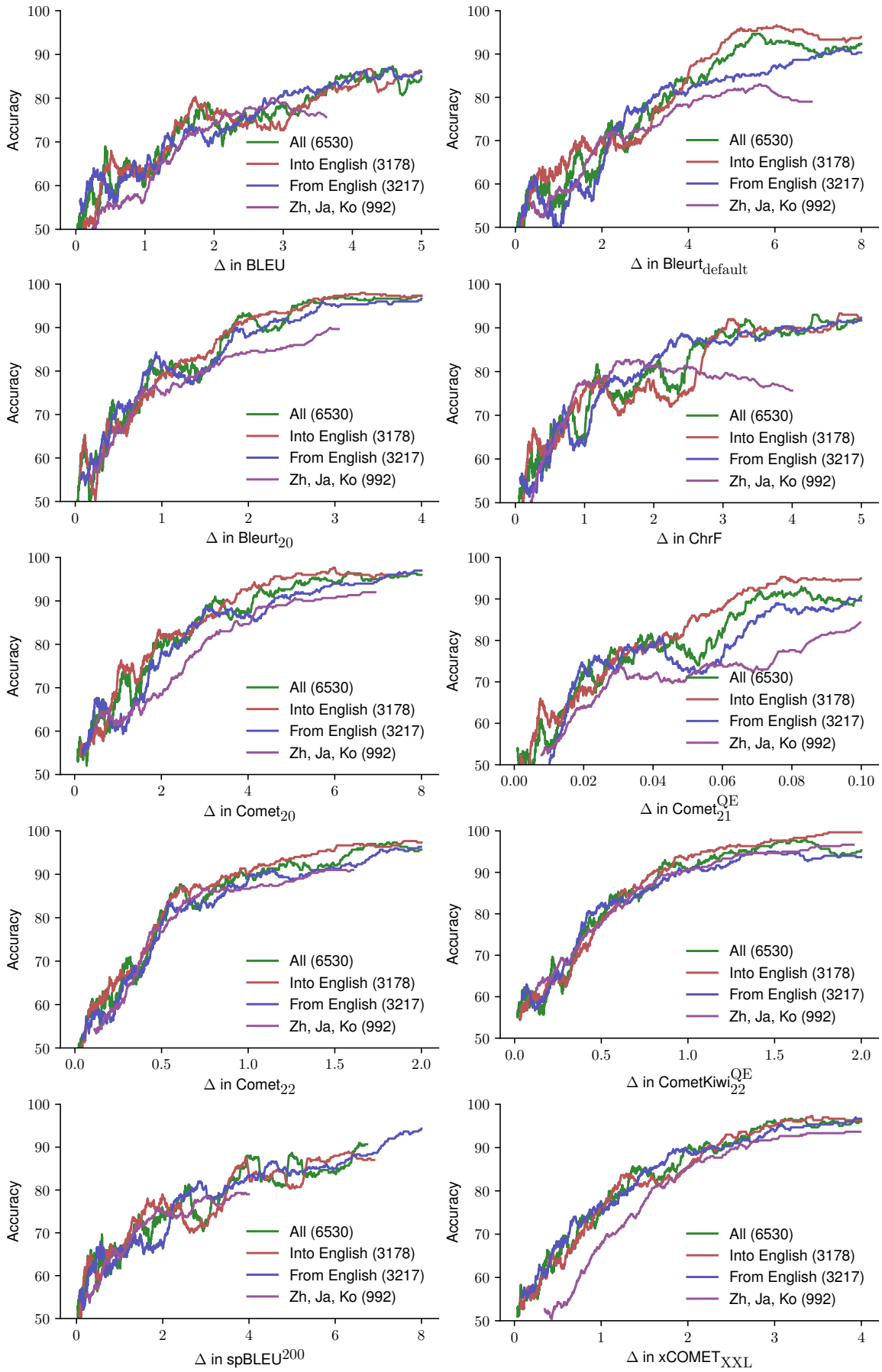

Figure 12: The comparison of pairwise accuracy on ToShip23 dataset when comparing into English, out-of-English, and Chinese, Japanese, Korean language pairs separately. The count shows total number of system-pairs in the evaluation. This figure extends Figure 6.

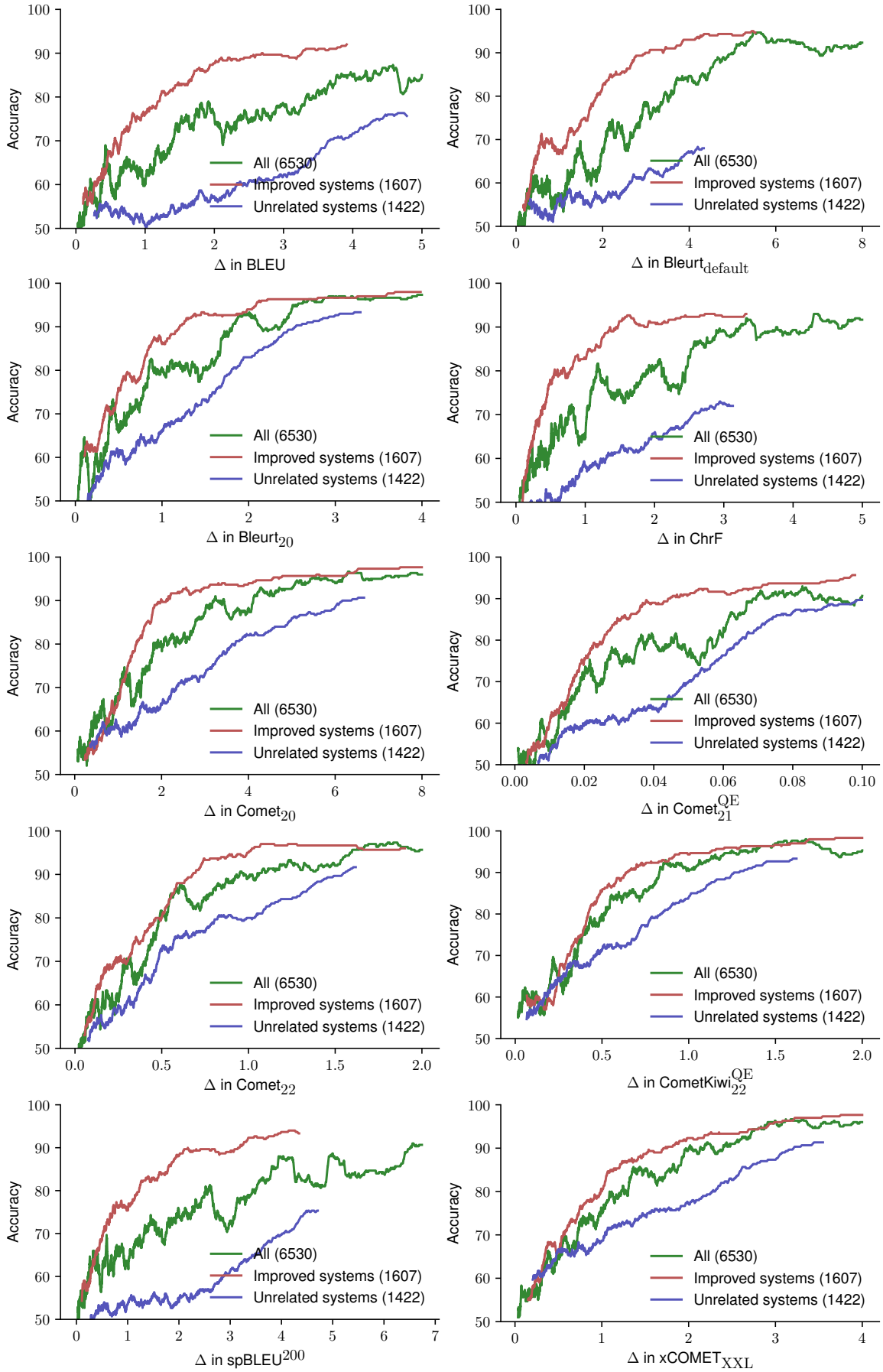

Figure 13: The comparison between iterated and unrelated systems on ToShip23. This figure extends Figure 7.