MT Metrics: What BLEU, chrF and COMET Do Not Measure

Jonathan Gosteli
Poly Research, Zürich, Switzerland
jonathan@poly-research.ch

September 2026

Abstract. Automatic metrics decide which machine translation systems are shipped, but the quantity they report—a scalar measure of similarity to a reference—is not the quantity a deployment decision needs, which is how often and how badly a system changes what a text means. We quantify the gap. On the WMT22 expert MQM annotations for zh→en, en→de and en→ru (63,328 rated segments, 15 systems per language pair) we measure the penalty BLEU and chrF assign to a translation as a function of the error a professional annotator actually found in it. A single major accuracy error costs 6.5 chrF relative to error-free translations of the same source segment; a single minor fluency or style error costs 2.1. The ratio is 3.2 for chrF and 1.7 for BLEU, against 6.2 in the human weighting these systems are ranked by, and in a third of cases the segment carrying the major error receives no penalty at all. On a controlled suite of 5,724 minimal pairs we find that deleting the word *not* costs less chrF than rewriting a date in an equally correct alternative format, and that giving a system a meaning-altering error in 10% of its segments costs it 0.30 BLEU—well inside the range in which reported differences are known to be uninformative. We argue that what is missing is severity rather than correlation, that COMET inherits the same deficiency from its regression target rather than from its architecture, and that reporting practice, not metric accuracy, is the tractable part of the problem.

1 Introduction

The standard workflow of machine translation research compresses a system's behaviour on a test set into one number and compares that number across systems. The number is well defined and cheap, and for the question it was designed to answer—which of two systems is closer, on average, to a set of reference translations—it works. The question that is actually asked of it in deployment is different: *how often does this system say something the source did not say, and how bad is it when it does?*

These are not the same question, and the difference is not a matter of correlation strength. A metric can correlate acceptably with human quality judgements at the system level while being nearly indifferent to the distinction between a translation that reads slightly awkwardly and a translation that inverts a negation, drops a clause, or names the wrong drug. That indifference is structural. BLEU [23] and chrF [24] count n-gram overlap; overlap is a function of how many tokens differ, not of which ones. COMET [27] replaces overlap with a learned regression, but regresses onto a scalar human score in which severity has already been collapsed into a weighted sum. In each case the output is a similarity, and severity is not recoverable from it.

The literature has established repeatedly that BLEU is a poor instrument [4, 16, 20, 29] and that neural metrics correlate better with human judgement [10, 11]. Our question is narrower and, we think, more actionable: given that the field has largely settled which metric correlates best, *what does the resulting number fail to tell you*, and by how much? We answer that empirically.

Our contributions are:

1. A severity-conditioned measurement of metric behaviour on real system output. Using expert MQM annotations [8, 19] as the ground truth for *which* error a segment contains, we measure the penalty BLEU and chrF assign to that segment relative to error-free translations of the same source, and show that the penalty is largely a function of edit size rather than of severity (§4).

2. A controlled minimal-pair suite of 5,724 pairs separating meaning-altering edits from meaning-preserving orthographic and formatting variation, showing that the two classes overlap heavily in metric space (§5).

3. A corpus-level detectability curve: how much of a system's output can be corrupted with meaning-altering errors before the corpus score moves by an amount anyone would notice (§6).

4. An analysis of what the corpus average hides across domains, and of what happens to the metric when it becomes the objective (§7).

5. An argument, with the published diagnostic evidence, that COMET does not repair the deficiency, and a concrete proposal for what to report instead (§§8–9).

2 What the three metrics compute

It is worth being precise about the object each metric returns, because the failure mode follows from it.

BLEU is the geometric mean of modified n-gram precisions for $n = 1 \ldots 4$ between a hypothesis and one or more references, multiplied by a brevity penalty. Every n-gram is worth the same. The metric is defined at the corpus level; the segment-level variant used here requires smoothing and is, by the original design's own terms, out of scope [29]. It is also a parameterised family rather than a single quantity, which is why scores are only comparable when the tokenisation and normalisation are pinned [26].

chrF [24] replaces word n-grams with character n-grams and combines precision and recall in an F_β score, standardly β=2 and character order 6. This makes it far more robust for morphologically rich targets and removes the tokenisation dependency, but it does not change the underlying accounting: the score is a count of matching character sequences. chrF++ [25] adds word unigrams and bigrams.

COMET [27, 28] encodes source, hypothesis and reference with a pretrained multilingual encoder and passes pooled representations to a feed-forward regressor trained to predict human quality scores. It is not a similarity count, and it is markedly better correlated with human judgement than either string metric. But its output is still one number on an uncalibrated scale, and the target it was trained to predict is a scalar summary of a human annotation in which a major error and several minor ones can be numerically identical.

All three are *reference-similarity* instruments. None of them is an error detector, and none of them was designed to be one.

3 Experimental setup

Data. We use the WMT22 General MT expert MQM annotations released by Google,[1] for zh→en, en→de and en→ru. Each segment was annotated by a professional linguist who marked error spans with a category and a severity (*major* or *minor*); the released segment score is the standard weighted sum, in which a major error counts 5 (25 for non-translation), a minor error 1, and a minor punctuation error 0.1 [8]. We join the span-level annotations to the released segment-score table by (system, document, hypothesis text), which is stable across the two files' differing segment-identifier conventions; quality-control items are excluded. The join covers 97.3%, 99.5% and 99.0% of annotated segments for zh→en, en→de and en→ru respectively, giving 63,328 scored segments over 15 systems per language pair. The annotated systems include a second human translation (*refB*, in zh→en and en→de) and systems produced by minimum Bayes risk decoding against BLEU, COMET and BLEURT; we exploit both below.

Metrics. BLEU, chrF and chrF++ are computed with SACREBLEU 2.6.0 [26]. Segment-level BLEU uses effective order with exponential smoothing.[2] Confidence intervals are bootstrap percentile intervals over 10,000 resamples. System-level comparisons are restricted to the segments annotated for every system in a language pair (1,150, 1,182 and 735 segments), so that all systems are scored on identical inputs.

A note on COMET. The environment in which these experiments were run has no network access to the model repositories that host COMET's weights, so COMET was not re-executed. Every number we report for BLEU, chrF and chrF++ is computed by us on the data described above; every claim we make about COMET's behaviour is either an argument from its construction or a citation to published diagnostic results, and is marked as such in §8. We regard this as a limitation of the study and flag it plainly rather than presenting borrowed numbers as our own.

4 The penalty a metric assigns does not track severity

If a metric is to be used as a proxy for translation quality, its penalty should scale with how bad the error is. We can test this directly, because the MQM annotations tell us exactly which error each segment contains.

Design. For each source segment we take the set of translations produced by the 15 systems and identify the subset that a professional annotator marked as containing *no* error. Against that error-free baseline we compare each translation of the same source segment that contains *exactly one* annotated error, grouped by that error's severity and by whether its category is an accuracy category (mistranslation, omission, addition, untranslated source, wrong named entity, hallucination, non-translation) or a form category (fluency, style, terminology, locale convention). Because the comparison is within a source segment, segment difficulty, length and reference quality are held constant. This yields 15,977 contrasts.

Result. Table 1 gives the outcome. The metrics do respond to severity, but weakly. Under chrF, a major accuracy error is penalised 3.2 times as heavily as a minor form error; under BLEU, 1.7 times. The human weighting the same segments are scored under separates them by a factor of 6.2. BLEU is barely able to tell the two apart: the interval for a major accuracy error, $[4.65, 5.57]$, is only 1.3 points clear of the interval for a minor style error, $[2.59, 3.37]$, and both are comfortably inside the range of variation that ordinary paraphrase induces.

More striking is the failure rate. In 33.3% of cases a segment that a professional annotator marked as containing a major accuracy error scores *at least as high* under chrF as the error-free translations of the same source segment; under BLEU the figure is 37.2%. Computed over all individual pairs rather than against the mean of the clean baseline, chrF

[1] https://github.com/google/wmt-mqm-human-evaluation

[2] Signatures, all with `nrefs:1`, `case:mixed`, `eff:yes`, `version:2.6.0`: BLEU `tok:13a`, `smooth:exp`; chrF `nc:6`, `nw:0`, `space:no`; chrF++ `nc:6`, `nw:2`, `space:no`.

Table 1: Metric penalty for a segment containing exactly one annotated error, relative to error-free translations of the same source segment. Pooled over zh→en, en→de and en→ru. Δ is the mean drop in the metric (higher = penalised more) with a 95% bootstrap interval. *No penalty* is the share of cases in which the erroneous segment scores at least as high as the error-free baseline under chrF; *inversion* is the same quantity computed over all individual error-free/erroneous pairs.

Human verdict on the segment	n	ΔBLEU	ΔchrF	ΔMQM	No penalty
one **major** accuracy error	4,984	5.11 [4.65, 5.57]	6.53 [6.08, 6.99]	5.08	33.3%
one **major** form error	597	3.30 [1.90, 4.65]	2.16 [1.10, 3.23]	4.99	42.2%
one **minor** accuracy error	3,053	3.48 [2.91, 4.06]	3.64 [3.13, 4.14]	0.99	41.4%
one **minor** form error	7,343	2.98 [2.59, 3.37]	2.06 [1.76, 2.35]	0.82	44.8%
ratio major accuracy : minor form		1.7×	3.2×	6.2×	

ranks the segment with the major accuracy error above an error-free translation of the same source in 35.2% of comparisons (41.6% for BLEU). These are cases in which the human verdict is not close: one translation carries a meaning-changing error and the other carries none.

Note also the ordering anomaly in the middle rows. A *minor* accuracy error is penalised more heavily by chrF (3.64) than a *major* form error (2.16), although the human weighting separates them in the opposite direction by a factor of five. The metric is not ranking by severity; it is ranking by how much surface material moved.

5 Minimal pairs: what costs more than a negation

The MQM experiment establishes that severity is weakly encoded, but it confounds severity with edit size: major errors are often larger edits. To separate them we built a controlled suite in which the edit is held small and only its *semantic* consequence varies.

Construction. From the zh→en data we take the 1,420 source segments for which at least one system produced a translation that the annotator marked error-free, and use that translation as the base hypothesis. We then apply two families of rule-based edits. The **meaning-altering** family instantiates the error types that critical-error detection targets [31]: deleting or inserting a negation, altering a number, substituting a named entity, swapping a gendered pronoun, substituting an antonym, deleting a clause, and appending a fluent unsupported sentence. The **meaning-preserving** family applies variation that a human annotator would score as no error or at most a locale convention issue: US/UK spelling, contraction expansion, numeral formatting, punctuation and quotation style, symbol and abbreviation expansion, date ordering, and deletion of the optional complementiser *that*. Each edit is applied only where applicable, giving 4,742 meaning-altering and 982 meaning-preserving pairs. Both members of a pair are scored against the same reference, so the quantity reported is the cost the metric charges for introducing that specific error into an otherwise good translation.

A stratified random sample of 140 pairs (10 per category) was inspected manually; 95% instantiate the intended category in fluent target-language text. The residual failures are mainly truncations that break a parenthetical.

Table 2: The minimal-pair suite. Δ is the mean penalty for introducing the edit; *no pen.* is the share of pairs in which chrF does not penalise the edited translation at all.

Edit	Class	n	ΔBLEU	ΔchrF	No pen.
clause omission	altering	1,299	7.48	12.04	2.2%
named entity substituted	altering	384	3.00	3.65	4.7%
date reformatted	**preserving**	30	**5.06**	**2.34**	6.7%
antonym substituted	altering	379	1.74	1.79	20.6%
hallucinated sentence	altering	1,420	5.36	1.78	25.6%
number altered	altering	291	3.53	1.52	10.3%
symbol expanded	**preserving**	38	**2.87**	**1.18**	23.7%
***that* deleted**	**preserving**	42	**1.51**	**1.13**	21.4%
negation removed / inserted	**altering**	**875**	**2.02**	**1.02**	**23.9%**
US/UK spelling	**preserving**	67	**1.57**	**0.97**	16.4%
gendered pronoun swapped	**altering**	**94**	**2.20**	**0.92**	**30.9%**
contraction expanded	preserving	279	2.13	0.69	45.5%
numeral reformatted	preserving	51	1.94	0.59	31.4%
punctuation restyled	preserving	475	0.26	−0.07	53.7%
all meaning-altering		4,742	4.67	4.57	16.0%
all meaning-preserving		982	1.27	0.43	43.7%

Result. In aggregate the metrics behave sensibly: meaning-altering edits are penalised roughly 3.7 (BLEU) to 10.6 (chrF) times as heavily as meaning-preserving ones. The aggregate is misleading. Table 2 and Figure 1 show that the two classes interleave. Deleting or inserting a negation—which inverts the proposition—costs 1.02 chrF. Rewriting *November 11* as *11 November*, which changes nothing, costs 2.34; expanding % to *percent* costs 1.18; switching *color* to *colour* costs 0.97. Swapping a gendered pronoun costs 0.92 chrF and is left entirely unpenalised in 31% of cases. Under BLEU the picture is no better and in places worse: expanding a contraction (2.13) is penalised more than inverting a negation (2.02), and reformatting a date (5.06) is penalised more than substituting a named entity (3.00) or altering a number (3.53).

Quantifying the overlap directly: if one draws a meaning-altering edit and a meaning-preserving edit at random, chrF assigns the smaller penalty to the meaning-altering one 25.9% of the time, and BLEU does so 31.4% of the time.[3] A

[3]Estimated over all 4,742 meaning-altering edits against 200 randomly drawn meaning-preserving edits each; ties counted as half.

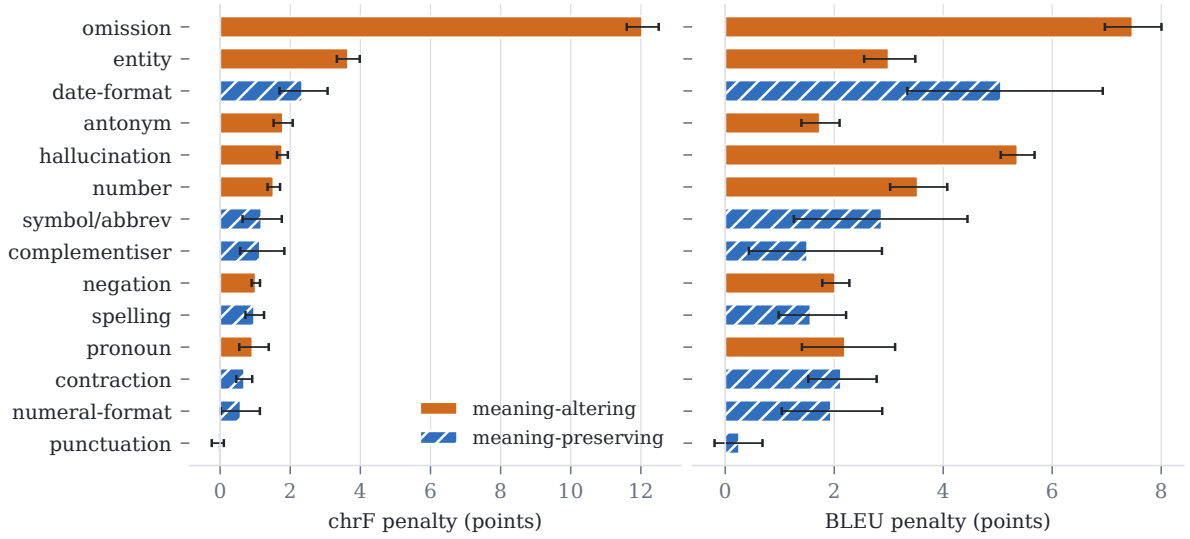

Figure 1: Mean metric penalty for introducing a single edit into an otherwise error-free translation, with 95% bootstrap intervals; categories ordered by chrF penalty. Meaning-altering and meaning-preserving edits overlap across most of the range, and the ordering is not stable between the two metrics.

metric used to decide whether a translation is safe would be wrong about which of two edits mattered roughly a third of the time.

Two of these categories deserve emphasis because they are the ones with identifiable downstream victims. Numbers and named entities are exactly where Amrhein and Sennrich [2] found COMET to be insensitive and where Alves et al. [1] found metrics generally to struggle; gender is the subject of a dedicated benchmark literature precisely because corpus metrics do not see it [30, 32].

6 At corpus level, meaning errors are close to invisible

The segment-level result becomes a deployment problem once scores are aggregated. We take the 1,420 base translations as a clean "system" (corpus BLEU 30.73, chrF 59.87), replace a random p fraction of its segments with a randomly chosen meaning-altering variant of that same segment, and recompute the corpus score, averaging over 30 draws.

Figure 2 shows the result. Corrupting 1% of segments costs 0.03 BLEU. Corrupting 10% costs 0.30 BLEU and 0.40 chrF. Corrupting 30%—a system that changes the meaning of nearly one segment in three—costs 0.90 BLEU and 1.20 chrF. It takes roughly a third of the corpus before the score moves by a single BLEU point, and Mathur et al. [20] show that differences of 1–2 BLEU correspond to genuine quality differences only about half the time, while Kocmi et al. [16] report a median BLEU difference of 1.3 points across 203 system pairs where BLEU flips a human-significant ranking. A meaning-error rate that would be unacceptable in any regulated setting is, in metric terms, indistinguishable from noise.

This is not an artefact of the corruption being subtle. The

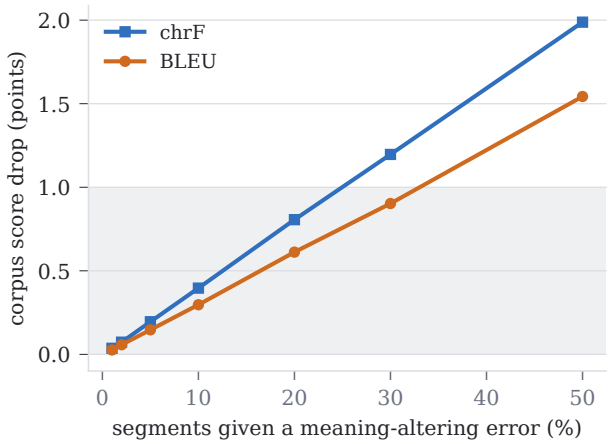

Figure 2: Corpus-level score drop as a function of the fraction of segments given a meaning-altering error. The shaded band marks drops below one point—the region in which reported metric differences have been shown to correspond to real quality differences only about half the time [20].

same edits are the ones a human annotator flags as major accuracy errors, and the same edits carry the entire risk profile of a translation system in medicine, law, or public safety.

7 What the average hides, and what the metric rewards

Domains. A corpus score is a mean, and a mean over heterogeneous material can be dominated by the easy part. Splitting the WMT22 data by its four domains and recomputing system-level rank correlation against the human MQM ranking within each domain (Table 3) shows how much the aggregate conceals. In zh→en, BLEU's agreement with hu-

man system ranking is $\tau = 0.58$ on news and $\tau = 0.03$ on social media; chrF's is $\tau = 0.56$ and $\tau = -0.03$. A single reported score is compatible with the metric being informative on one part of the distribution and uninformative—or anti-informative—on another. Across the twelve language-pair/domain cells, the system that BLEU ranks first is the system human annotators rank first in only three.

The reference is not the best translation. The WMT22 release includes a second human translation, *refB*, scored by the same annotators as an ordinary system. In en→de, expert annotators rank it 2nd of 15; BLEU ranks it 11th and chrF 13th (Table 4). A metric that scores a professional human translation below ten machine systems is not measuring translation quality; it is measuring proximity to one particular translator's choices—a point made a decade of work ago [7] and still operative.

The metric as objective. The same release includes systems produced by minimum Bayes risk decoding against each of four metrics [9, 10], which lets us watch Goodhart's law directly. In en→ru the system that optimised BLEU is ranked 1st of 15 by BLEU and 1st by chrF, but only 2nd by human annotators, behind a system BLEU places 8th. In zh→en the system that optimised BLEURT is ranked 3rd of 15 by human annotators and 12th by BLEU; the system that optimised COMET is 7th by humans and 14th by BLEU. Optimising a neural metric moves output away from the reference's surface form, and the string metrics record that movement as a loss of nine and seven rank positions respectively—a loss the humans do not see. Whichever metric is used as the objective ceases, at that moment, to be usable as the evaluation [5].

8 Why COMET does not solve this

The obvious rejoinder is that these are string metrics, and the field has moved on. COMET and its relatives correlate far better with human judgement than BLEU does: the WMT22 metrics task ranked BLEU last of thirteen reference-based metrics and recommended abandoning it [10], and WMT24 placed BLEU, spBLEU and chrF 23rd, 22nd and 20th of 26 [11]. On correlation, this is not in dispute.

The deficiency we have measured is not correlation, and it is not repaired by a better encoder, for three reasons.

First, the target. COMET is trained to regress a scalar. Whether that scalar is a direct assessment or an MQM score, severity has already been collapsed into it by a fixed weighting before the model sees it. A model that predicted MQM perfectly would still be unable to distinguish a segment with one major accuracy error from a segment with five minor errors, because MQM assigns them the same number. Under the standard weighting, that identification is exact by construction. The information a deployment decision needs is destroyed upstream of the metric, in the definition of the label.

Second, the empirical record. Diagnostic suites built specifically to test this find the same blind spots we find. Karpinska et al. [14], over 31K minimal pairs, report that metrics including COMET "struggle to differentiate the severity of some critical errors, such as an addition of a plausible but meaning-changing word. . . or incorrect number," and that COMET-QE is more sensitive to word repetition than to named-entity replacement or sentence negation. Amrhein and Sennrich [2] show COMET models are "not sensitive enough to discrepancies in numbers and named entities," and that the bias is "hard to fully remove by simply training on additional synthetic data"—a caveat worth reading against Juraska et al. [13], whose authors report augmenting training data specifically to patch "fluent but unrelated translation" and "undertranslation." Amrhein et al. [3] and Moghe et al. [22], benchmarking 47 metrics on 68 accuracy error types, find no clear winner and that "most metrics ignore the source sentence" and "tend to prefer surface level overlap." In the most recent campaign, Lavie et al. [18] report that where language, dialect and script mismatches were the critical issue, "the automatic metrics are unable to detect the catastrophic translations."

Third, interpretability and comparability. Even where the score is accurate, it is not readable. Zouhar et al. [34] document that "COMET scores are not comparable between papers or even technical setups," and propose a signature scheme analogous to SACREBLEU's. Kocmi et al. [17] examine how large a score difference must be before humans notice, across translation direction, domain and system closeness, and find the requirement depends sharply on the last: $+2$ BLEU between iterated versions of one model agrees with humans about 90% of the time, while the same $+2$ BLEU between unrelated systems is "barely better than toss of a coin." A fixed threshold therefore does not transfer between settings. And Moghe et al. [21] find that chrF, COMET and BERTScore all show "negligible correlation" with downstream task outcomes.

The constructive reading of this literature is that the field already knows the answer, and it is not a better scalar. It is metrics that emit error spans and severities rather than a number—xCOMET [12], AutoMQM [6], GEMBA-MQM [15]—which the WMT24 results show can be done without sacrificing correlation, xCOMET having placed in the top cluster [11]. Our contribution here is to say how much is lost by not doing so, in the units practitioners actually report.

9 What to report instead

We do not propose a new metric. We propose that a translation quality report contain the following, and note that all of it is already computable with existing tooling.

1. **A severity-partitioned error rate, not only a similarity score.** The primary reported quantity should be the rate of major accuracy errors, with a confidence interval, on a sample large enough to bound it. Our results show that this rate is not recoverable from BLEU, chrF, or any scalar; it has to be estimated by an error-span metric or by annotation. Reporting a corpus score alongside it is fine; reporting it alone is not.

Table 3: System-level Kendall τ against expert MQM, computed within each domain over 15 systems, alongside the pooled figure. The pooled number is not a summary of the per-domain numbers so much as a blend of a working regime and a failing one.

	Metric	conversation	e-commerce	news	social	*pooled*
zh→en	BLEU	0.54	0.33	0.58	**0.03**	0.35
	chrF	0.54	0.39	0.56	**−0.03**	0.47
en→de	BLEU	0.49	0.24	0.14	0.37	0.35
	chrF	0.47	0.35	0.24	0.37	0.31
en→ru	BLEU	0.41	0.49	0.47	0.64	0.56
	chrF	0.45	0.43	0.54	0.64	0.64

Table 4: Selected systems, ranked out of 15 within each language pair on the segments annotated for every system. ∗*bestmbr* systems were produced by minimum Bayes risk decoding against the named metric; *refB* is a second human translation.

	System	MQM	rank	BLEU	rank	chrF	rank
en→de	Online-W	−0.76	1	37.27	5	64.54	5
	refB (human)	−0.93	2	33.66	11	61.48	13
	bleu_bestmbr	−0.95	3	39.59	1	65.00	2
	comet_bestmbr	−1.04	6	33.56	12	61.93	11
en→ru	Online-W	−1.49	1	32.53	8	58.51	7
	bleu_bestmbr	−1.98	2	35.79	1	59.86	1
	comet_bestmbr	−2.29	6	28.76	13	55.77	13
zh→en	refB (human)	−1.74	1	47.17	1	68.82	1
	bleurt_bestmbr	−2.62	3	26.75	12	57.59	12
	comet_bestmbr	−2.80	7	22.73	14	52.30	14

2. **Disaggregation by the axes that matter.** Report per-domain and, where relevant, per-speaker-group or per-dialect scores. Table 3 shows that the pooled figure can average a working regime with a failing one.

3. **Contrastive accuracy on the known blind spots.** Numbers, named entities, negation, omission and addition, and gender should be reported as accuracies on a purpose-built contrastive set [1, 3, 14, 32]. These are cheap to construct—the suite in §5 took a few hundred lines of rules—and they measure what corpus scores cannot.

4. **Pinned configurations and no cross-paper comparison.** Report SACREBLEU signatures [26] and the COMET equivalent [34]; do not compare scores across papers, setups, or language pairs [17].

5. **Never the objective as the evaluation.** If a metric was used in decoding, reranking, filtering or training, it is disqualified as the evaluation of that system [2, 5, 33].

6. **Human MQM on a subsample as the anchor.** Expert MQM with document context [8] remains the reference standard; a few hundred annotated segments bound the major error rate far better than a corpus score over thousands.

10 Limitations

COMET was not re-run in this study, for the reason given in §3; our COMET claims are argumentative and citational, not experimental, and a replication that scores the same minimal-pair suite with COMET, CometKiwi, xCOMET and MetricX would sharpen or refute them. The minimal-pair suite is rule-based and English-target only; the meaning-preserving categories, while individually defensible, have small sample sizes in three cases ($n \leq 51$) and their confidence intervals are correspondingly wide. Our "meaning-preserving" label is our own judgement, validated by inspection but not by independent annotation; a strict annotator might score a reformatted date as a minor locale convention error, which would make the comparison with negation more favourable to the metrics but would not close the gap. The MQM data covers three language pairs and four domains from a single evaluation campaign, and the annotations are single-rater. Finally, the corruption experiment in §6 inserts errors uniformly at random; a real system's errors are correlated with segment difficulty, which would change the shape of the curve though not, we expect, its order of magnitude.

11 Conclusion

BLEU, chrF and COMET measure similarity to a reference translation. Used for the purpose they were built for—ranking systems by average closeness to a reference—they work, imperfectly and with well-documented caveats. Used as proxies for whether a system is safe to deploy, they answer a question nobody asked. On real annotated system output, a metric penalty separates a meaning-changing error from a stylistic blemish by a factor of three where the human standard separates them by a factor of six, and fails to penalise the meaning-changing error at all a third of the time. On controlled minimal pairs, deleting a negation costs less than changing a date format. At corpus level, a system can change the meaning of a tenth of its output for the price of a third of a BLEU point.

None of this is an argument for a new number. It is an argument that the number is the wrong shape. The severity distinction that deployment decisions turn on is destroyed at the point where an error annotation is collapsed into a scalar, and no amount of correlation improvement downstream of that point can recover it. The metrics that emit error spans and

severities already exist and already perform competitively. What is missing is the reporting convention that requires them.

Reproducibility. All experiments use publicly released data and SACREBLEU with the signatures given in §3. The minimal-pair suite, the join procedure, and the analysis scripts accompany this note.

References

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

[2] Chantal Amrhein and Rico Sennrich. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the ACL and the 12th International Joint Conference on Natural Language Processing*, pages 1125–1141, 2022.

[3] Chantal Amrhein, Nikita Moghe, and Liane Guillou. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation*, pages 479–513, 2022.

[4] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 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, 2006.

[5] Daniel Deutsch, Rotem Dror, and Dan Roth. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, 2022.

[6] Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 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, 2023.

[7] Markus Freitag, David Grangier, and Isaac Caswell. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 61–71, 2020.

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

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

[10] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 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*, pages 46–68, 2022.

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

[12] Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024.

[13] Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, 2024.

[14] Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, 2022.

[15] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, 2023.

[16] Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 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, 2021.

[17] Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1999–2014, 2024.

[18] Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, 2025.

[19] Arle Richard Lommel, Hans Uszkoreit, and Aljoscha Burchardt. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, (12):455–463, 2014.

[20] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 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, 2020.

[21] Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13060–13078, 2023.

[22] Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137, 2025.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

[24] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, 2015.

[25] Maja Popović. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, 2017.

[26] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, 2018.

[27] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702, 2020.

[28] 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. COMET-22: Unbabel-IST 2022 submission for

the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585, 2022.

[29] Ehud Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, 2018.

[30] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874, 2021.

[31] Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, 2021.

[32] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, 2019.

[33] Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 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*, pages 5428–5443, 2023.

[34] Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, 2024.