

An Analysis of Source-Side Diversity in Machine Translation Evaluation

Esther Ploeger¹, Johannes Bjerva², Dong Nguyen¹, Robert Östling³

¹Utrecht University ²Aalborg University ³Stockholm University

Correspondence: e.ploeger@uu.nl

Abstract

Benchmarks are central to assessing progress in machine translation (MT), yet the content of widely used MT test sets is increasingly scrutinized. Concerns about data contamination and benchmark difficulty have gained traction, but the role of source-side diversity remains largely overlooked. Many popular benchmarks contain overlapping source items and even exact duplicates, raising the question of how such repetition affects evaluation quality. We empirically examine how source-side diversity shapes MT evaluation. First we measure the inter-source diversity of nine popular MT datasets. Next, our downstream analysis, based on WMT24, suggests that benchmarks with low diversity may be less discriminative and may provide artificially narrow confidence intervals. Ultimately, we call for greater attention to inter-source diversity in general MT benchmark design.

1 Introduction

Designing general-purpose machine translation (MT) benchmarks is fundamentally difficult: They are expected to reflect the diversity of real-world translation while avoiding an unintended focus on any single style, domain, or register. Recent efforts in WMT evaluation campaigns have prioritized this, aiming to evaluate “MT systems’ ability to handle a broad range of translation and language use” and “explore the translation capabilities of current systems in diverse settings” (Kocmi et al., 2024). Still, MT benchmarks contain source-side items with considerable overlap: Figure 1 shows such an example of two items in the WMT24 benchmark. This overlap suggests potential correlation between items: A model that performs well on item #346 would also be expected to perform well on item #350. Yet, how this affects the quality of MT evaluation at large is currently an open question.

Previous work (Freitag et al., 2020; Bawden et al., 2020; Zeng et al., 2024) has mainly addressed

src #346:	Aaaand it's still offline .
ref A:	Uuuuund immer noch offline .
GPT-4:	Uuund es ist immer noch offline .
src #350:	Still offline .
ref A:	Immer noch offline .
GPT-4:	Immer noch offline .

Figure 1: Example of **overlapping** source-side items (#346 and #350) in WMT24 (`en` → `de`). GPT-4 translated items similarly, leading to repeated output trigrams and similar scores across items.

MT benchmark diversity from the *reference-side*, aiming to mitigate the monotonous style of typical references used in common MT benchmarks. In this work, we instead investigate *source-side* diversity. We address the following research question: **How does source-side diversity affect MT evaluation?**

More specifically, we focus on *inter-source diversity*, based on pairwise comparisons of source items. In particular, our analysis targets lexical, surface-form repetition between items. Intuitively, if a benchmark contains two identical items, the weight of this item in a final aggregate metric is doubled, compared to a unique item. This can lead to (unintended) skews in the linguistic phenomena that are represented. This is problematic when the goal is to draw conclusions about translation quality generally, and results are aggregated uniformly across items, as is common practice. Furthermore, running large-scale MT evaluation campaigns requires human translations, human annotations of quality, and inference over all competing models. When repeated and similar items are included, such resources are potentially wasted: (Near-)duplicates

provide limited new information about model performance. Despite these intuitive arguments, there has been no systematic investigation of the quantitative, downstream impact of inter-source diversity in MT benchmarks. We therefore present the following contributions:

- An **audit** of source-side diversity in nine popular MT benchmarks (§3);
- An **empirical analysis** of the downstream impact of controlled inter-source diversity in WMT24, regarding model rankings, confidence intervals, and discriminativeness (§5);
- **Further analysis** concerning item and score similarity, domains, and the relationship between difficulty and diversity (§6);
- A **discussion** on the role of inter-source diversity in MT evaluation (§7);
- Practical **recommendations** for MT benchmark design (§8).

Our results, based on WMT24, indicate that source-side diversity is an important and beneficial property of MT benchmarks. We highlight the need for more systematic attention to inter-source diversity in general-purpose MT benchmark creation, and provide a practical algorithm for controlling benchmark diversity in existing benchmarks.

2 Background & Related Work

Redesigning MT Benchmarks. Benchmarks have played a vital role in the automatic evaluation of MT developments, and thereby steering research directions. This relationship is bidirectional: Benchmarks shape progress, but progress also reshapes what we require from strong benchmarks. Currently, a prominent criticism of existing MT benchmarks concerns their *difficulty*: As the state of the art improves, benchmarks risk becoming too easy, reducing their ability to meaningfully differentiate system performance. Mitigating such benchmark saturation has become a popular line of MT research. Zhan et al. (2021) and Liu et al. (2022) approached difficulty from the perspective of *metrics*, where different benchmark items get different weights in aggregate scores, based on their predicted difficulty. Data-level approaches pursued difficulty by filtering for specific textual characteristics, such as selecting longer sentences (Chen et al., 2023). A recent line of work has extended this direction. Proietti et al. (2025) propose the task

of predicting the difficulty of benchmark items, and use these predictors to sample difficult items, constructing more challenging benchmarks. Xu et al. (2025) search for difficult test items on the internet, and Pombal et al. (2025) and Zouhar et al. (2026) generate items that are difficult to translate (the former from scratch; the latter iteratively). This line of research has practical impact: The WMT25 general MT benchmark was explicitly filtered for difficulty (Kocmi et al., 2025).

Difficult benchmarks are not necessarily linguistically representative, as they aim to assess MT performance specifically in challenging scenarios. In fact, Zouhar et al. (2026) demonstrate a trade-off: Optimizing solely for difficulty may come at the expense of diversity. Since most efforts to increase benchmark difficulty have rarely accounted for diversity, source-side diversity has been largely absent from the discourse on improving MT evaluation. Our work addresses this perspective directly.

Diversity in MT Benchmarks. Most works that address diversity in MT benchmarks focus on the reference-side. Freitag et al. (2020) note that WMT references exhibit a monotonous style, and include additional, manually paraphrased references, which increased the correlation with human judgment on the WMT19 English→German set. Bawden et al. (2020) found that, when using automatically generated paraphrases, improvements (based on BLEU) were small and inconsistent. Zeng et al. (2024) experimented with using LLM-generated reference alternatives on four translation directions in WMT22, and report an improved correlation with human judgment. These works define the diversity of a set of references according to lexical diversity (as opposed to e.g. syntactic diversity), because MT models and many metrics are sensitive to lexical, surface-level variation.

We similarly focus on lexical diversity, but approach diversity from the *source*-side. This offers a practical advantage: Source-side interventions could lead to more efficient use of evaluation resources. That is, more informed selection of source-side items can avoid ineffective efforts such as translating identical (or highly similar) items, both in terms of human effort as well as model inference costs. Since benchmarks are commonly translated into a range of languages (9 for our WMT24 source set, 55 languages in WMT24++; Deutsch et al., 2025), this effort is not negligible.

3 An Audit of Source-Side Diversity

Understanding whether source-side overlap meaningfully affects MT evaluation first requires establishing how common such overlap actually is. If the examples in Figure 1 prove to be isolated cases, and overlap is generally rare in widely used MT benchmarks, then inter-source diversity may pose little practical concern. However, the current state of source-side diversity in MT evaluation sets remains largely undocumented. In this section, we therefore conduct a systematic audit of source-side diversity across popular MT benchmarks: What is the extent of duplicates and overlap present in benchmark items?

Datasets. We analyze nine of the most widely used MT evaluation datasets: Europarl (Koehn, 2005), FloRes-101/200 devtest (Goyal et al., 2022; NLLB Team et al., 2024), BOUQuET (Andrews et al., 2025), and six recent benchmarks from the main WMT evaluation campaigns (WMT20-25; Barrault et al. 2020; Akhbardeh et al. 2021; Kocmi et al. 2022, 2023, 2024, 2025). This broad selection includes the most recent evaluation datasets used in the OPUS-MT Dashboard (Tiedemann and de Gibert, 2023). We analyze English sources. For the WMT datasets, we take the source text that was translated into most target languages that year. Details for each dataset are found in Appendix A.

Diversity Measures. We are principally interested in *inter-source diversity*, i.e. how similar or dissimilar individual test instances are to one another. To assess this, we first count the number of items that occur more than once in a benchmark (# Dpl), and the number of near-duplicates

(# N-Dpl) as the number of items in a test set that have a Levenshtein similarity of > 0.9 to another item in the test set.¹ Note that this is a conservative threshold, capturing only items with very high surface-form similarity, excluding examples such as in Figure 1 (Levenshtein similarity: 0.65). Additionally, we measure lexical overlap through the average Dice-coefficient of the character bigrams (cf. Kondrak, 2005) over all pairs of items, and subtract this from 1 to retrieve an average lexical distance.² This mean pairwise distance (MPD) measures how lexically different benchmark items are from each other, on average. Additionally, we report the average distance (again: $1 - \text{Dice-coefficient}$) to each item’s nearest neighbour (mean minimal pairwise distance; MMPD). While MPD captures the global lexical spread of the benchmark, MMPD reveals to what extent items are locally clustered.

We also examine how inter-source diversity relates to other *overall diversity* measures, commonly used in MT (e.g. Vanmassenhove et al., 2019). We first measure the number of unique words in each benchmark ($|W|$). This measure naturally increases as the number of items in a benchmark increases. Therefore, we also calculate the length-robust MTLT (average length of consecutive word strings that maintain a certain TTR value; McCarthy and Jarvis, 2010).³

¹We use TheFuzz (<https://github.com/seatgeek/thefuzz>), using the method `fuzz.ratio()`. This threshold was argued in previous work to account for minor variations (e.g. Szep et al., 2026).

²We experimented with a range of other metrics, including Levenshtein Distance, Jaro-Winkler Distance and Hamming Distance, but found that overall trends remain the same.

³We calculate this using https://github.com/kristopherkyle/lexical_diversity, default settings.

Benchmark	# Items	Inter-source Diversity				Overall Diversity	
		# Dpl (↓)	# N-Dpl (↓)	MPD (↑)	MMPD (↑)	W (↑)	MTLD (↑)
Europarl	3,064	3	17	0.59	0.39	8,336	102.61
FloRes	1,012	0	0	0.62	0.45	5,818	108.26
BOUQuET	854	2	2	0.75	0.51	2,957	106.20
WMT20	1,418	0	0	0.61	0.42	8,583	125.36
WMT21	1,002	0	0	0.63	0.44	5,724	118.81
WMT22	2,037	204	249	0.72	0.40	6,107	97.30
WMT23	2,074	6	20	0.73	0.47	6,480	92.57
WMT24	998	9	17	0.72	0.46	6,885	120.68
WMT25	332	0	0	0.45	0.33	7,173	114.52

Table 1: Audit of source-side diversity in nine widely used MT benchmarks. Arrows indicate directions which suggest higher diversity. Single values indicating **most** and **least** diversity are marked in each column.

Results and Discussion. Table 1 shows the results. Firstly, we observe that the majority (5/9) of benchmarks we analyzed contain exact duplicates (# Dpl) at the source-side. Although the number of items that appear more than once is generally not very high (e.g. 6 out of 2,074, i.e. 0.3% in WMT23), our audit reveals WMT22 as a problematic outlier, with more than 200 repeated items. Surface-level investigation shows that several items appear frequently in that benchmark (e.g. “Go to your Home screen.”, 11 times; “Tap Settings”, 11 times; “Tap Device information”, 7 times).

Overall, mean pairwise distances (MPD) are below 0.80, suggesting that benchmark items are not all lexically well-separated. Given that MT evaluation is typically budget-constrained, higher MPD values would be desirable, as they indicate a greater coverage of distinct linguistic phenomena. The mean minimum pairwise distance (MMPD) shows that, apart from BOUQuET, all benchmarks have mean nearest-neighbour distances below 50%. This means that, for most items, another source sentence in the same benchmark is somewhat lexically similar, revealing substantial local overlap despite moderate global diversity.

Lastly, we observe that *how* diversity is measured matters: Overall diversity scores, which are widely used in MT (e.g. Vanmassenhove et al., 2019), and inter-source diversity results vary. For example, WMT20 contains the highest number of unique words and retrieves the highest MTLT, but its mean pairwise distance (MPD) between items is lower than for example BOUQuET, which retrieves the highest MPD of 0.75, with a more modest MTLT score of 106.20.

Implications. This audit indicates that source-side overlap is not an isolated issue, but a recurring characteristic in popular MT benchmarks, making it worthwhile to investigate these patterns empirically in relation to MT evaluation. We find that source-side diversity manifests in more subtle ways than mere exact duplication: The absence of exact duplicates does not necessarily imply higher overall diversity scores, or high mean pairwise distance. Practically, our results suggest that inter-source diversity cannot be taken for granted in existing MT benchmarks, and that trivial solutions (e.g. applying a de-duplication filter) do not necessarily mitigate the issue.

4 Experimental Setup

To evaluate the downstream impact of inter-source diversity, we first construct subsets with controlled levels of inter-source similarity (§4.2). Based on these subsets, we evaluate differences in model rankings, confidence intervals and discriminativeness (§4.3).

4.1 Data

We base our analysis on the WMT24 benchmark (Kocmi et al., 2024). This dataset includes model- and item-level scores across a range of recent MT models and LLMs. Importantly, it is not an outlier in terms of diversity scores (§3), unlike for example WMT22 (# Dupl) and WMT25 (MPD), and it was not pre-filtered for difficulty, unlike WMT25. WMT24 has also shaped several recent claims about state-of-the-art MT, for example “*The LLM Era Is Here but MT Is Not Solved Yet*” (Kocmi et al., 2024), and continues to be used actively beyond the original shared task (e.g. Zouhar et al., 2026; Ramos et al., 2026). Moreover, it serves as the foundation for the broader WMT24++ dataset (Deutsch et al., 2025), further underscoring its relevance. For our analysis, we use the English sources which were translated into nine target languages (Czech, German, Spanish, Icelandic, Japanese, Russian, Ukrainian, Chinese, Hindi). We include each of these language directions in our analysis, to strengthen the cross-lingual generalizability of our findings, and to demonstrate robustness against reference-side variation (references differ across target languages).

4.2 Controlling Benchmark Diversity

Dataset filtering has proven to be an effective approach for mitigating redundancy in NLP benchmarks more generally (Rubinstein et al., 2026; Gupta et al., 2025; Gevers et al., 2025) and for redesigning MT benchmarks for difficulty (Proietti et al., 2025). We extend this approach to controlling diversity, by sampling high and low diversity benchmarks.

Sampling Algorithm. To retrieve a diverse sample, in which overlap is minimized, we approach the problem as an instance of the *Maximum Minimum Diversity Problem* (MMDP), where the objective is to select items such that the minimum distance between any two sampled items is maximal. This eliminates the need to specify any manual diversity threshold. Since the MMDP is known

Benchmark	# Items	Inter-source Diversity				Overall Diversity	
		# Dpl ($\downarrow$)	# N-Dpl ($\downarrow$)	MPD ($\uparrow$)	MMPD ($\uparrow$)	W ($\uparrow$)	MTLD ($\uparrow$)
MAXMIN	499	0	0	0.81	0.59	2,416	258.80
MINMIN	499	0	0	0.52	0.38	6,024	138.52
Random*	499.0	2.40	4.20	0.72	0.48	4,463.60	170.92
<i>st. dev</i>	± 0.0	± 2.27	± 2.57	± 0.01	± 0.01	± 112.75	± 7.26

Table 2: Diversity evaluation on WMT24 subsets, where Random, MAXMIN and MINMIN sample sizes are 50% of the original dataset. *For all random samples in this paper, we report the average over 10 runs.

to be NP-hard we apply a heuristic; greedy sampling. Given a set of source items S , for each $s \in S$, we first calculate each distance $dist$ with respect to every other source $s' \in S \setminus \{s\}$. For this, we use $1 -$ the average Dice-coefficient of the character bigrams, consistent with §3. Given this matrix of pairwise distances, we then aim to sample k items with maximum minimum distance. We first identify the source $s \in S$ that has the maximum distance $dist$ with respect to all other sources $s' \in S \setminus \{s\}$. Then, until k is reached, we retrieve the next point that has the largest minimum distance to all items in the current sample. A formal representation of our sampling approach is found in Appendix B. We sample using a threshold of 50% of the full dataset⁴, resulting in a MAXMIN set which we expect to exhibit increased lexical diversity compared to random samples of the same size. As a controlled low diversity benchmark, we construct a sample using the opposite objective (minimum minimum distance): MINMIN.

Diversity Verification. To verify whether our algorithm effectively controls source-side diversity, we compare the constructed MAXMIN and MINMIN test sets using the diversity measures from §3. For a length-fair comparison with the full test set, we report the average over ten random samples of the same size (50%). Table 2 shows that, indeed, our controlled MAXMIN sample yields substantially increased lexical diversity, across most measures. On the other hand, diversity in our MINMIN dataset is effectively decreased. The only exception, the number of unique words, is explained by the MINMIN sample containing more longer items (note that $|W|$ is very sensitive to length). Further, neither MAXMIN nor MINMIN contains duplicates. MAXMIN and MINMIN overlap by 68 items (13.6%). As a final verification step, we visu-

⁴The 50% threshold is arbitrary; we use it to demonstrate the difference between high and low diversity benchmarks, rather than searching for an optimal value.

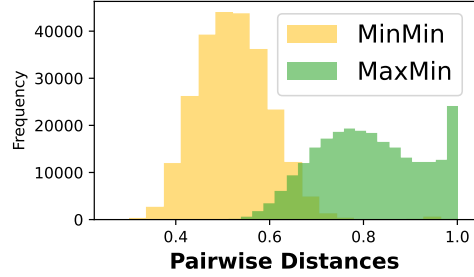


Figure 2: Frequency distribution of pairwise distances (1 – Dice bigram coefficients) across the MINMIN and MAXMIN samples.

alize the distribution of pairwise distances in Figure 2: MAXMIN indeed contains more dissimilar items than MINMIN.

4.3 Measuring Downstream Impact

Unlike model evaluation, where performance can be measured directly on a benchmark, evaluating a benchmark itself is more challenging. In previous work, benchmark filtering approaches (e.g. Gupta et al. 2025; Zouhar et al. 2025) are deemed successful if the performance of the filtered set corresponds to that of the full benchmark. However, because existing benchmarks themselves are not guaranteed to be diverse (§3), we cannot rely on the full set as a gold standard. Instead, we first inspect how **model rankings** differ across the MAXMIN, MINMIN and random samples (§5.1). We then look at differences in terms of their **confidence intervals** (§5.2) and **discriminativeness** (§5.3). We report chrF in the paper’s main text, which was also used in the OCELoT leaderboard for WMT24 (Kocmi et al., 2024). We rely on an automatic metric because human assessment (e.g. MQM, ESA) is not available for each item for each model. We choose chrF instead of a learned metric (e.g. COMET) in the main text, since learned metrics may disadvantage low-resource languages (Falcão et al., 2024).⁵

⁵Appendices D, E and G show results on MT metrics beyond chrF; The observed patterns hold.

	cs	de	es	hi	is	ja	ru	uk	zh
All $\leftrightarrow$ MAXMIN	0.98	0.98	0.99	0.97	0.98	0.95	0.99	0.97	0.99
All $\leftrightarrow$ MINMIN	0.94	0.96	0.94	0.98	0.99	0.97	0.92	0.96	0.98
All $\leftrightarrow$ Random*	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99
<i>st. dev</i>	± 0.005	± 0.011	± 0.008	± 0.004	± 0.006	± 0.008	± 0.007	± 0.005	± 0.005

Table 3: Correlation (Spearman’s ρ) of model rankings, when calculating average performance per model based on 50% subsets of the benchmark items (MAXMIN, MINMIN, Random). All results are statistically significant ($p < 0.001$).

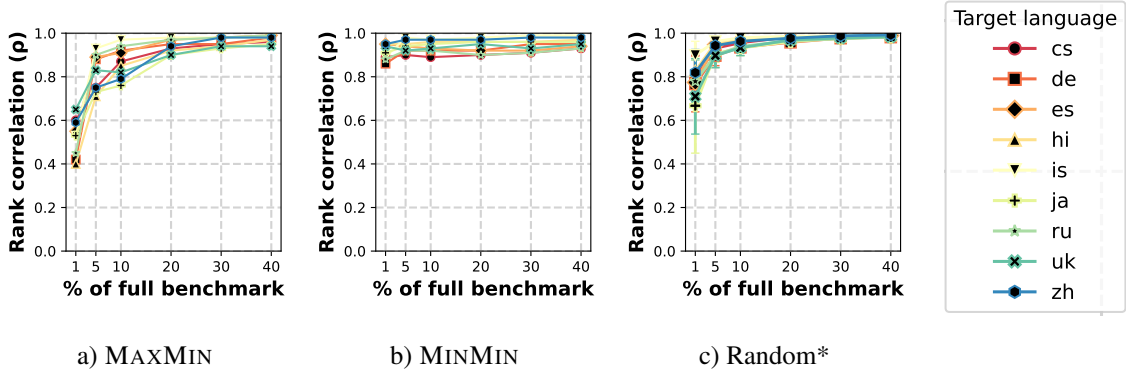


Figure 3: Rank correlations with the full benchmark (Spearman’s ρ) across controlled diversity samples (MAXMIN, a; MINMIN, b) and Random* samples (c) of different sizes: 1%, 5%, 10%, 20%, 30%, 40% of WMT24 sources.

5 Results and Analysis

5.1 Model Rankings

Table 3 shows the correlation between model rankings when calculating average performance per model based on only subsets of the benchmark items (MAXMIN, MINMIN and 10 random samples). All correlations are strong (> 0.90) and statistically significant; model rankings stay virtually the same, whether evaluating on MAXMIN, MINMIN, a random sample, or the full benchmark. While similar model-level conclusions can be drawn with less data, perhaps WMT24 is not diverse enough to show a meaningful difference.

While differences in correlation between MAXMIN and MINMIN (50% of the full benchmark, i.e. 499 lines) are marginal in Table 3, there is a stark difference when analyzing smaller sample sizes. Figure 3 shows the rank correlation when taking controlled diversity samples of different sizes (1%, 5%, 10%, 20%, 30%, 40% of the full WMT24 benchmark). Taking only 1% (9 sentences) items through minimum minimum diversity sampling yields a strong (and statistically significant) correlation (> 0.85) with the full benchmark, across all nine translation directions. This strong correlation is consistent across sampling sizes. By contrast, sampling for maximum minimum diver-

sity requires more sentences to be included, for a strong correlation with the full benchmark. Note that correspondence to the full test set is not necessarily a good indicator of benchmark quality, since the full benchmark may include overlap (See §3 and Figure 1). We therefore evaluate benchmark behavior more generally in the following sections.

5.2 Performance and Confidence Intervals

Benchmarks with many highly similar items potentially produce artificially narrow confidence intervals, because resampling-based estimates (such as bootstrap intervals) implicitly assume that each item contributes independent information, while this is not the case. For example, [Hardt \(2026\)](#) writes that copying each benchmark data point 10 times will make the confidence intervals overly optimistic: While no new information on model performance is added, the apparent statistical certainty increases, purely as an artifact of this duplication.

Figure 4 shows the 95% bootstrap confidence intervals for the MAXMIN and MINMIN datasets. The benchmark with lower diversity exhibits a consistently narrower interval, demonstrating exactly this effect: Reduced variety in the evaluation set leads to deceptively tight confidence intervals, potentially overstating how reliably the benchmark shows model performance. The figure also shows

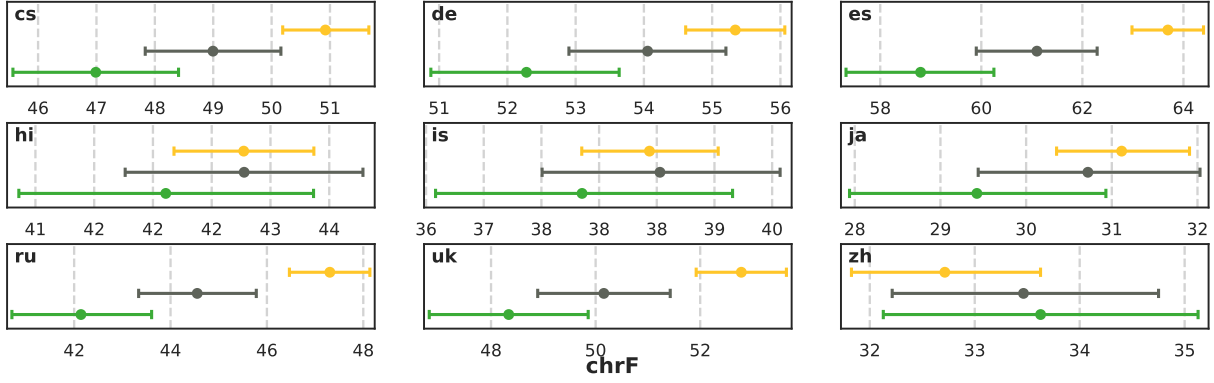


Figure 4: 95% bootstrap confidence intervals (sampled 10,000 times with replacement) for the **MAXMIN** (bottom) and **MINMIN** (top) and **Random*** (10 runs) (middle) samples.

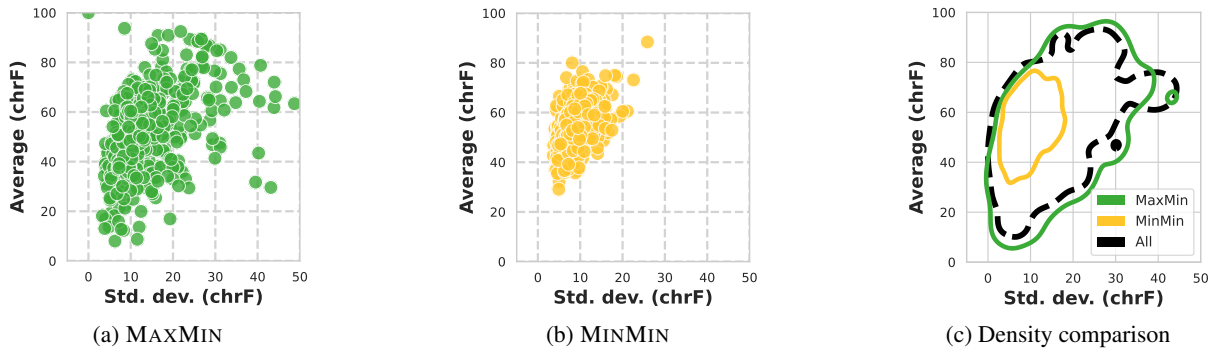


Figure 5: Distribution of average chrF scores and their standard deviation across benchmark items in MAXMIN (a), MINMIN (b), and a comparison with the full set of benchmark items (c), for WMT24 en→de.

that the less diverse benchmark’s scores are generally higher. This indicates that low inter-source diversity items may have inflated overall chrF scores for WMT24.

5.3 Discriminativeness

A requirement for an informative benchmark is that it includes items on which models perform differently, that is, different models obtain different scores. Inspired by dataset cartography from training dynamics (Swayamdipta et al., 2020), we plot each item in a (sampled) benchmark according to its average performance over all models (chrF, *y*-axis) and its standard deviation across model performance (chrF, *x*-axis). Comparing Figure 5a (MAXMIN) and Figure 5b (MINMIN) shows a stark contrast: Benchmark items in the less diverse controlled benchmark cluster together more than in the more diverse benchmark, both on the *x*-axis and *y*-axis. In Figure 5c we compare this with the full benchmark: The spread of the (smaller, 50%) MAXMIN sample resembles that of the full set, while MINMIN results are more dense. We quantify these differences in Table 4, where we also

compare with the average over 10 random samples again. On the *y*-axis (Δ Avg.), the smaller range for MINMIN means that average performance is more similar. Thus, this benchmark is less informative when it comes to mapping MT models’ strengths and weaknesses. On the *x*-axis (Δ Std. Dev.), this means that the performance differences *between* models are smaller. Thus, the less diverse benchmark has weaker discriminative capabilities than the more diverse one. This holds across all 9 target languages (Appendix F).

	Δ Std. Dev.	Δ Avg.	Avg. Coord. Distance
MAXMIN	48.72	92.00	20.79
MINMIN	22.24	59.28	10.50
Random*	44.88	88.00	17.35
<i>st. dev</i>	± 2.26	± 4.36	± 0.63

Table 4: Difference between min and max value on the *x*-axis (Δ Std. Dev.), *y*-axis (Δ Avg.) and the average coordinate distance for each item in the sampled benchmarks, for WMT24 en→de. *Results over the random samples are averaged over 10 runs.

	cs	de	es	hi	is	ja	ru	uk	zh
Intercept	21.3*	21.90*	24.38*	20.72*	23.30*	22.09*	22.89*	17.54	17.67
Lex. Sim.	-27.19*	-26.54*	-30.36*	-23.19*	-30.88*	-34.03*	-26.25*	-26.79*	-23.10*
Syn. Sim.	-3.81	-1.35	-1.89	-6.53	-3.63	-7.26	-6.70	-0.76	-6.79
Sem. Sim.	1.36	0.09	-0.70	-0.94	-1.99	3.86	0.92	6.42*	5.62*

Table 5: Results for multiple regression on distance matrices for similarity dimensions and average chrF score differences in WMT24. Statistically significant ($\alpha < 0.01$) correlations (1,000 permutations) are marked with an asterisk. Coefficients are rounded to two decimals. Results for COMET, MetricX-24 and BLEU are in Appendix H.

6 Further Analysis

The previous empirical analysis indicates that inter-source diversity meaningfully affects MT evaluation: While model rankings change minimally, the low-diversity sampled benchmark produces more narrow confidence intervals and is less discriminative, compared to a benchmark sample with higher diversity. This merits further investigation: We first verify the assumption that similar sources tend to retrieve similar scores (§6.1), analyze the sampled domains (§6.2), and further investigate the relationship between difficulty and diversity (§6.3).

6.1 Similar Sources, Similar Scores?

The intuition that inter-source diversity affects MT evaluation relies on the assumption that evaluating on similar items tends to yield similar outputs, thus resulting in similar scores. In this section, we verify this assumption, and explore three dimensions of similarity (lexical, syntactic and semantic).

Methodology. To assess to what extent similar items retrieve similar results, we correlate pairwise (lexical, syntactic and semantic) similarity matrices (over all pairs of source items) with the pairwise average score difference matrix (over all pairs of item-level results). We report results with chrF in the main text (see Appendix H for other MT metrics). We run multiple regression on distance matrices (MRM) (Legendre et al., 1994; Lichstein, 2007), with all three item similarity matrices as predictors. A negative correlation coefficient can be interpreted as follows: The higher the similarity between two test items, the smaller the distance in their average MT score across models.

Similar to the main experiments, for *lexical* similarity, we take the Dice-coefficient of the character bigrams in two sources. Using part-of-speech (POS) tags is an established approach for calculating “syntactic footprints” (Nerbonne and Wiersma, 2006). To approximate syntactic distance, we first

obtain the POS tags using Stanza (Qi et al., 2020). We then compute the Jaccard similarity by dividing the number of overlapping POS-bigrams in the two sources by their union. More complex approaches, such as tree kernel distances, are prohibitively computationally expensive for all pairwise combinations. To approximate *semantic* similarity, we calculate the commonly used cosine similarity (e.g. Guo et al., 2024) between sentence embeddings. We use Sentence-BERT (Reimers and Gurevych, 2019) to obtain embeddings. Specifically, we use all-mpnet-base-v2, as the providers claim that this model “provides the best quality”.⁶ Formal definitions for these similarity measures are in Appendix C.

Results and Discussion. Table 5 shows a consistent, negative, statistically significant weak to moderate correlation between lexical similarity and (average) score differences of items. That is, a high similarity between two test items is associated with a low difference in their output scores; Lexically similar items tend to give similar results. Although there are many factors that influence automatic MT metrics (such as variation in reference formulations), it is interesting to see such a consistent result in the lexical dimension, while the syntactic and semantic similarity do not show such patterns. This supports our focus on *lexical* diversity in our main analyses. The results are mostly robust across reference variation, since the pattern holds across all target languages (each with their own references). Similar conclusions hold for BLEU, COMET-22 and MetricX-24 scores (Appendix H).

6.2 Domain-Level Sample Comparison

In the WMT24 evaluation campaign, benchmark items were sampled from four domains in an attempt to increase the diversity of evaluated phe-

⁶https://sbert.net/docs/sentence_transformer/pretrained_models.html

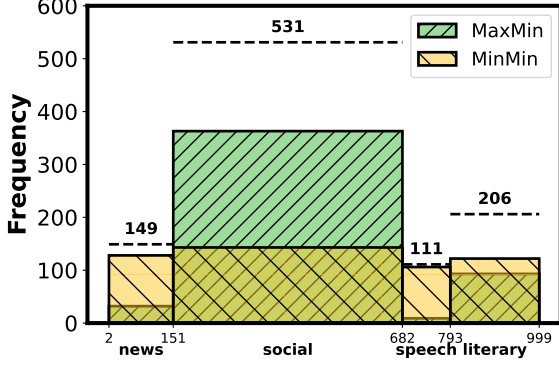


Figure 6: Number of items sampled per domain in MAXMIN and MINMIN, with total items per domain indicated by horizontal dashed lines. Note that the total size is 997 instead of 998, because the canary sentence (item #1) does not belong to a domain.

nomena (Kocmi et al., 2024). We investigate how MAXMIN and MINMIN compare in terms of their distribution across these domains, and what this implies about benchmark creation.

Figure 6 shows that the social domain is sampled more in MAXMIN, compared to MINMIN. In MINMIN, almost all items from the news and speech domain are included in the sample. In Table 6 we investigate this further. The news and speech domains exhibit a considerably lower MPD (0.53 and 0.47, respectively) than the other two domains (0.79 and 0.68). Interestingly, social contains the only (near-) duplicates in WMT24, but still retrieves the highest MTLD, MPD and MMPD. High MPD and MMPD may in part be explained by item length: Although our lexical distance metric corrects for length, shorter items inherently contain fewer bigrams, and therefore tend to exhibit lower overlap than longer items. Regardless, this suggests that simply sampling from *more* domains may be less effective for increasing inter-source diversity than sampling from more *diverse* domains.

6.3 Difficulty vs. Diversity

Considering the recent interest in increasing the difficulty of MT benchmarks (§2), it is relevant to investigate how diversity and difficulty interact. The results in §5 showed that chrF scores for the less diverse test set (MINMIN) were generally higher than those in Random and MAXMIN, indicating a lower level of difficulty, for WMT24. Inspired by Zouhar et al. (2026), we now look at it the other way around: the diversity of benchmarks sampled for difficulty. In MT, difficult source-side items are

Domain	# Dpl	MPD	MMPD	W	MTLD
news	0	0.53	0.39	2,715	132.72
social	9	0.79	0.50	2,456	134.41
speech	0	0.47	0.37	2,152	98.81
literary	0	0.68	0.46	2,219	120.41

Table 6: Diversity audit per WMT24 domain.

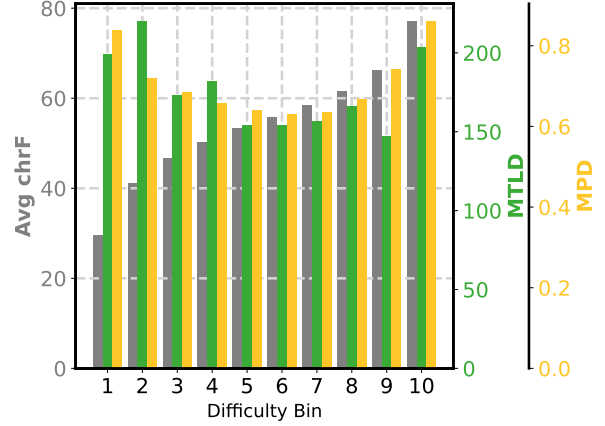


Figure 7: Source-side diversity scores across bins of decreasing difficulty in WMT24, en→de.

commonly defined as items for which MT models retrieve low scores (Proietti et al., 2025). Therefore, we operationalize an item’s difficulty through its overall score across models.

We sort all benchmark items by average performance (en→de, chrF) and categorize the items into 10 bins of decreasing difficulty. For each of these bins, we calculate the MPD and MTLD scores. Figure 7 shows the results, including average chrF scores per bin. Both in terms of inter-source diversity (MPD) and overall lexical diversity (MTLD), diversity is highest for the easiest and most difficult bins, while diversity scores are generally lower for the mid-difficulty bins. One explanation for this is that some highly diverse items are tricky to translate given a single reference, because of non-standard language use, for example: “@user47 nooooo0000oo0o0ooooo 🤪”. When it comes to items from the easy bins, addresses or hyperlinks may be frequently seen during training, and thus easy for MT models, but can still inflate lexical diversity scores, for example: “The Tierra del Sol Gallery is located at 7414 Santa Monica Blvd. For information, visit tierradelsolgallery.org.” This suggests that the relation between difficulty and diversity may be more nuanced than a strict trade-off (Zouhar et al., 2026).

7 Discussion

This paper has presented an empirical analysis of source-side diversity in MT evaluation. At the same time, determining what belongs in a benchmark is not purely an empirical question; it can be argued from multiple perspectives and therefore calls for a broader, more conceptual discussion.

A common intuition is that general-purpose benchmarks “should reflect the full range of linguistic variation [...] that is used in the relevant domain, context, and language variety” (Bowman and Dahl, 2021). In MT specifically, diversity is often linked to generalizability: Goyal et al. (2022) note that a benchmark that only consists of very short and straightforward sentences may be “not as generalizable to a more diverse set of content”. Similarly, *general* WMT evaluation campaigns have explicitly aimed to include “diverse settings” (Kocmi et al., 2024), reinforcing the idea that broad linguistic coverage is desirable for assessing general MT capabilities.

However, the opposite perspective also deserves explicit consideration. First, it should be noted that it is not by default problematic to evaluate MT models on low-diversity text. Domain-specific applications, such as the automatic translation of instruction manuals, are expected to be evaluated in these low-diversity domains. However, the issue that we address in this paper is that of suggested generalizability in combination with unawareness regarding the inter-source diversity. Claims about “general MT capabilities” (Kocmi et al., 2022) or overall MT architectures (‘*The LLM Era is Here...*’; WMT24 (Kocmi et al., 2024)) implicitly assume that the evaluation set is sufficiently diverse to support such conclusions. A second consideration is that low inter-source diversity may be a natural consequence of document-level MT evaluation. When benchmark items are drawn from coherent documents, they will often share topic and style, making the data more similar than benchmarks with independently sampled sentences. This helps explain why the document-level WMT25 benchmark yields lower MPD than benchmarks constructed through sentence-level sampling (§3). Such overlap may be desirable for coherent document-level evaluation, but it still poses challenges when results are interpreted as evidence of general MT performance, when metrics are aggregated over overlapping (and correlated) items, or statistical tests assume independence on the sentence-level.

8 Conclusion and Recommendations

Many general-purpose MT benchmarks contain source-side overlap (§3), but whether and how this can impact MT evaluation downstream was still an open question. We addressed this research question through systematically controlling source-side diversity for English source items in WMT24. We found that model rankings were relatively stable across samples, so model comparisons were not substantially affected. Still, the low-diversity benchmark produced narrower confidence intervals and was shown to be less discriminative, compared to the higher-diversity benchmark.

Recommendations. Based on our findings, we offer several practical recommendations for future MT benchmark development and evaluation. First, we recommend *inter-source* diversity to be taken into account more explicitly when creating new benchmarks. At minimum, applying a deduplication filter on the source-side could avoid unexpected repeated measurements and save evaluation resources. We encourage researchers to report diversity scores when introducing or using MT benchmarks, enabling more transparent insights into *what* is measured, and how this compares to other benchmarks. While model rankings are relatively stable across diversity samples, we caution against drawing absolute conclusions about the state of MT based on benchmarks that exhibit limited diversity, as such datasets may not reflect the full range of linguistic phenomena that modern MT systems encounter, potentially leading to overfitting on specific language varieties. Finally, we advise against evaluating (with uniformly aggregated metrics) on benchmarks with known diversity issues, notably WMT22.

Inter-source diversity has so far received limited attention in MT evaluation research, and our findings highlight several directions for further investigation. A first question is whether the effects we observe in WMT24 generalize to other evaluation campaigns, and how this evolves as MT models continue to improve. Future work could also explore diversity sampling strategies in benchmark creation beyond greedy selection. Lastly, how to *aggregate* results over overlapping or correlated items remains an open methodological problem, especially for already existing benchmarks.

Limitations

Datasets. Our downstream analysis is conducted on the English sources in the WMT24 benchmark. There are multiple reasons for this: WMT24 spans multiple domains, was not filtered for difficulty (as WMT25), is not an outlier in terms of number of duplicates (Table 1), and is still commonly used for evaluating MT models (e.g. in WMT24++). We have included results across metrics (chrF, Popović, 2015; COMET-22, Rei et al., 2022; MetricX-24, Juraska et al., 2024) and all target languages, to increase generalizability. While evaluating on only one WMT test set is not uncommon in MT evaluation research (e.g. Zouhar et al. (2026)), we cannot (and do not) claim that the diversity results would hold to all (present and future) MT benchmarks.

Focus On Subsampling. Improving benchmarks through subsampling is common practice (Rubinstein et al., 2026; Gupta et al., 2025; Gevers et al., 2025; Proietti et al., 2025; Zouhar et al., 2025). This allows us to control the diversity of the sample directly. We leave the generation of new (diverse) test items (cf. e.g. Zouhar et al. (2026)) in a controlled manner to future work.

Broader Benchmark Issues. Gevers et al. (2025) write the following about improving NLP benchmarks: “it is important not to overestimate the usability of single solutions, since they are often created with only one or a few issues in mind, neglecting other pitfalls.” While source-side diversity is a commonly overlooked property of MT benchmarks, benchmark quality is multifaceted. Approaches to diverse and difficult benchmarks are not necessarily mutually exclusive. We note that issues in major MT benchmarks are not limited to diversity. For example, in our surface-level qualitative analysis, we observed anonymized usernames (“@user”, WMT24), HTML code (“<div id=sec1></div>”, WMT24) and other templates (“Question No 4 by (H-0820 / 00)”, Europarl). These are not necessarily representative of real-world MT usage (e.g. it can be expected that few users would like to translate HTML code from English to Spanish (and German, and Hindi, and Icelandic, etc.)). As a result, such ‘general’ benchmarks do not necessarily reflect how well such models perform on actual end users’ tasks. All in all, this highlights the need to pay (even) closer attention to evaluation datasets in MT.

Acknowledgments

We would like to thank the anonymous reviewers, as well as the members of the AAU-NLP group at Aalborg University and the ICS-NLP group at Utrecht University, for their helpful feedback on earlier drafts of this paper. We are especially grateful to Heather Lent and Anna Wegmann for providing detailed comments on previous versions of this work. This work is supported by the ERC Starting Grant Data Divers 101162980.

References

- Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, and 17 others. 2021. [Findings of the 2021 conference on machine translation \(WMT21\)](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.
- Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathaniel Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenco, Albert Ventayol-Boada, and Shireen Yates. 2025. [BOUQuET : dataset, benchmark and open initiative for universal quality evaluation in translation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.
- Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joannis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, and 2 others. 2020. [Findings of the 2020 conference on machine translation \(WMT20\)](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.
- Rachel Bawden, Biao Zhang, Lisa Yankovskaya, Andre Tättar, and Matt Post. 2020. [A study in improving BLEU reference coverage with diverse automatic paraphrasing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 918–932, Online. Association for Computational Linguistics.

- Samuel R. Bowman and George Dahl. 2021. [What will it take to fix benchmarking in natural language understanding?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.
- Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang, Zongyao Li, Jiaxin Guo, Ning Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. [Multifaceted challenge set for evaluating machine translation performance](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 217–223, Singapore. Association for Computational Linguistics.
- Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. [WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.
- Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. [COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565, Torino, Italia. ELRA and ICCL.
- Markus Freitag, David Grangier, and Isaac Caswell. 2020. [BLEU might be guilty but references are not innocent](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.
- Ine Gevers, Victor De Marez, Jens Van Nooten, Jens Lemmens, Andriy Kosar, Ehsan Lotfi, Nikolay Banar, Pieter Fizez, Luna De Bruyne, and Walter Daelemans. 2025. [In benchmarks we trust ... or not?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23662–23676, Suzhou, China. 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.
- Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. [The curious decline of linguistic diversity: Training language models on synthetic text](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.
- Vipul Gupta, Candace Ross, David Pantoja, Rebecca J. Passonneau, Megan Ung, and Adina Williams. 2025. [Improving model evaluation using SMART filtering of benchmark datasets](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4595–4615, Albuquerque, New Mexico. Association for Computational Linguistics.
- Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc’Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english.
- Moritz Hardt. 2026. *The Emerging Science of Machine Learning Benchmarks*. Princeton University Press.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. [MetricX-24: The Google submission to the WMT 2024 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouagna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. 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, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. [Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. 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, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023.

- Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.
- Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. [Findings of the 2022 conference on machine translation \(WMT22\)](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Philipp Koehn. 2005. [Europarl: A parallel corpus for statistical machine translation](#). In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.
- Grzegorz Kondrak. 2005. N-gram similarity and distance. In *International symposium on string processing and information retrieval*, pages 115–126. Springer.
- Pierre Legendre, François-Joseph Lapointe, and Philippe Casgrain. 1994. Modeling brain evolution from behavior: a permutational regression approach. *Evolution*, 48(5):1487–1499.
- Jeremy W Lichstein. 2007. Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecology*, 188(2):117–131.
- Yilun Liu, Shimin Tao, Chang Su, Min Zhang, Yanqing Zhao, and Hao Yang. 2022. [Part represents whole: Improving the evaluation of machine translation system using entropy enhanced metrics](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*, pages 296–307, Online only. Association for Computational Linguistics.
- Philip M McCarthy and Scott Jarvis. 2010. Mtd, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.
- John Nerbonne and Wybo Wiersma. 2006. [A measure of aggregate syntactic distance](#). In *Proceedings of the Workshop on Linguistic Distances*, pages 82–90, Sydney, Australia. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, 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, and 20 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- José Pombal, Nuno M Guerreiro, Ricardo Rei, and Andre Martins. 2025. [Zero-shot benchmarking: A framework for flexible and scalable automatic evaluation of language models](#). In *Second Conference on Language Modeling*.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. [Estimating machine translation difficulty](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24261–24285, Suzhou, China. Association for Computational Linguistics.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Miguel Moura Ramos, Duarte M Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M Guerreiro, Ricardo Rei, Nicolas Boizard, and 1 others. 2026. Eurollm-22b: Technical report. *arXiv preprint arXiv:2602.05879*.
- 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.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Alexander Rubinstein, Benjamin Raible, Martin Gubri, and Seong Joon Oh. 2026. [Disco: Diversifying sample condensation for efficient model evaluation](#). In *International Conference on Learning Representations*, volume 2026, pages 15229–15255.
- Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. [Dataset cartography: Mapping and diagnosing datasets with training dynamics](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

- Marton Szep, Jorge Marin Ruiz, Georgios Kaissis, Paulina Seidl, Rüdiger von Eisenhart-Rothe, Florian Hinterwimmer, and Daniel Rueckert. 2026. [Unintended memorization of sensitive information in fine-tuned language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6461–6480, Rabat, Morocco. Association for Computational Linguistics.
- The Omnilingual MT Team, Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Eric Saint-James, Ioannis Tsiamas, Chierh Cheng, Joe Chuang, Paul-Ambroise Duquenne, Mark Duppenhaler, Nate Ekberg, Cynthia Gao, Pere Lluís Huguet Cabot, João Maria Janeiro, Jean Mailard, and 12 others. 2026. [Omnilingual MT: Machine translation for 1,600 languages](#).
- Jörg Tiedemann and Ona de Gibert. 2023. [The OPUS-MT dashboard – a toolkit for a systematic evaluation of open machine translation models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–327, Toronto, Canada. Association for Computational Linguistics.
- Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. [Lost in translation: Loss and decay of linguistic richness in machine translation](#). In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.
- Wenda Xu, Vilém Zouhar, Parker Riley, Mara Finkelstein, Markus Freitag, and Daniel Deutsch. 2025. Searching for difficult-to-translate test examples at scale. *arXiv preprint arXiv:2509.26619*.
- Xianfeng Zeng, Yijin Liu, Fandong Meng, and Jie Zhou. 2024. [Towards multiple references era – addressing data leakage and limited reference diversity in machine translation evaluation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11939–11951, Bangkok, Thailand. Association for Computational Linguistics.
- Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021. [Difficulty-aware machine translation evaluation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 26–32, Online. Association for Computational Linguistics.
- Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025. How to select datapoints for efficient human evaluation of nlg models? *Transactions of the Association for Computational Linguistics*, 13:1789–1811.
- Vilém Zouhar, Wenda Xu, Parker Riley, Juraj Juraska, Mara Finkelstein, Markus Freitag, and Daniel Deutsch. 2026. [Generating difficult-to-translate texts](#).

In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 204–219, Rabat, Morocco. Association for Computational Linguistics.

A Dataset Details

Table 7 (next page) contains an overview of the evaluation datasets analyzed in this work, including paper references, download links, licenses and additional information. For the WMT datasets, the listed download links are no longer available; still, data can be found in the following repository: <https://github.com/google-research/mt-metrics-eval>.

B Sampling Algorithm

Algorithm 1 MaxMin Diversity Sampling.

```

1:  $s \leftarrow \arg \max_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \text{dist}(s, s')$ 
2:  $M \leftarrow \{s\}$ 
3: while  $|M| < k$  do
4:    $M \leftarrow M \cup \{\arg \max_{s \in \mathcal{S} \setminus M} (\min_{s' \in M} \text{dist}(s, s'))\}$ 
5: end while
6: return  $M$ 

```

C Calculation of Similarity Metrics

Lexical Similarity.

$$\text{sim}_{lex}(s, s') = \frac{2 * |N(s) \cap N(s')|}{|N(s)| + |N(s')|} \quad (1)$$

where $N(s)$ denotes the set of n-grams in source sequence s , and $N(s')$ denotes the set of n-grams in source sequence s' :

Syntactic Similarity.

$$\text{sim}_{syn}(s, s') = \frac{|N(pos_s) \cap N(pos_{s'})|}{|N(pos_s) \cup N(pos_{s'})|} \quad (2)$$

where $N(pos_s)$ denotes the set of POS n-grams in source sequence s , and $N(pos_{s'})$ denotes the set of POS n-grams in source sequence s' :

Semantic Similarity.

$$\text{sim}_{sem}(s, s') = \frac{E(s) \cdot E(s')}{|E(s)| |E(s')|} \quad (3)$$

where the semantic similarity between two sources s and s' is defined as the cosine similarity between the embeddings of s and s' , given embedding model E .

D Model Ranking Results Across Metrics

Tables 8 and 9 show the results of the model rank analysis (§5.1) when using COMET-22 and MetricX-24, respectively, instead of chrF. The same conclusion holds: Model rankings are relatively stable across samples.

E Discriminativess Results Across Metrics

Tables 10 and 11 show the results of the model discriminativeness analysis (§5.3) when using COMET-22 and MetricX-24, respectively, instead of chrF. We observe similar patterns as in the main text.

F Discriminativess Analysis Across Target languages

Tables 12–19 contain the results for the analysis of benchmark discriminativeness (§5.3), across the eight other target languages. Results and conclusions are consistent with those based on en→de in the main analyses.

G Confidence Interval Results Across Metrics

Figure 8 and 9 show the results of the confidence interval analysis (§5.2) when using COMET-22 and MetricX, instead of chrF.

H MRM test results with COMET-22, MetricX-24 and BLEU

Table 20 shows the MRM correlation results with COMET-22 as an automatic evaluation metric. The coefficients are smaller, but note that COMET scores are on a different scale than BLEU scores. Table 22 shows the MRM correlation results with MetricX-24 as an automatic evaluation metric. Table 21 shows the MRM correlation results with BLEU as an automatic evaluation metric.

Dataset	Reference(s), data source, license, notes
FloRes-101/200 (devtest)	Paper references: Guzmán et al. (2019); Goyal et al. (2022) Source: https://dl.fbaipublicfiles.com/flores101/dataset/flores101_dataset.tar.gz Dataset license: Creative Commons Attribution-ShareAlike 4.0
Europarl	Paper reference: Koehn (2005) Source: https://www.statmt.org/wmt06/shared-task/test2006.en.gz Dataset license: Not disclosed explicitly
BOUQuET (test)	Paper reference: Andrews et al. (2025); Team et al. (2026) Source: https://huggingface.co/datasets/facebook/bouquet Dataset license: Meta Dataset Research License Agreement Note: We take the test set on the sentence_level.
WMT20	Paper reference: Barrault et al. (2020) Source: https://storage.googleapis.com/mt-metrics-eval/mt-metrics-eval-v2.tgz Dataset license: Apache-2.0 license Note: We take the English source file which was used for the translation directions: <i>en</i> → <i>cs</i> , <i>de</i> , <i>es</i> .
WMT21	Paper reference: Akhbardeh et al. (2021) Source: https://storage.googleapis.com/mt-metrics-eval/mt-metrics-eval-v2.tgz Dataset license: Apache-2.0 license Note: We take the English source file which was used for the translation directions: <i>en</i> → <i>cs</i> , <i>zh</i> , <i>en</i> , <i>de</i> .
WMT22	Paper reference: Kocmi et al. (2022) Source: https://storage.googleapis.com/mt-metrics-eval/mt-metrics-eval-v2.tgz Dataset license: Apache-2.0 license Note: We take the English source file which was used for the translation directions: <i>en</i> → <i>zh</i> , <i>uk</i> , <i>ja</i> , <i>de</i> , <i>cs</i> , <i>ru</i> .
WMT23	Paper reference: Kocmi et al. (2023) Source: https://storage.googleapis.com/mt-metrics-eval/mt-metrics-eval-v2.tgz Dataset license: Apache-2.0 license Note: We take the English source file which was used for the translation directions: <i>en</i> → <i>zh</i> , <i>cs</i> , <i>uk</i> , <i>ru</i> , <i>ja</i> , <i>he</i> .
WMT24	Paper reference: Kocmi et al. (2024) Source: https://storage.googleapis.com/mt-metrics-eval/mt-metrics-eval-v2.tgz Dataset license: Apache-2.0 license Note: We take the English source file which was used for the translation directions: <i>en</i> → <i>cs</i> , <i>de</i> , <i>es</i> , <i>is</i> , <i>ja</i> , <i>ru</i> , <i>uk</i> , <i>zh</i> , <i>hi</i> .
WMT25	Paper reference: Kocmi et al. (2025) Source: https://storage.googleapis.com/mt-metrics-eval/mt-metrics-eval-v2.tgz Dataset license: Apache-2.0 license Note: We take the English source file which was used for all translation directions.

Table 7: Additional information about the evaluation datasets that were analyzed in this work.

	cs	de	es	hi	is	ja	ru	uk	zh
All ↔ MAXMIN	0.95	0.99	0.98	0.99	1.00	0.98	0.99	0.94	0.99
All ↔ MINMIN	0.97	0.99	0.94	0.99	0.99	0.96	0.97	0.95	0.95
All ↔ Random*	0.99	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99
<i>st. dev</i>	± 0.005	± 0.003	± 0.008	± 0.003	± 0.00	± 0.006	± 0.005	± 0.013	± 0.005

Table 8: Correlation (Spearman’s ρ) of model rankings, when calculating average performance (**COMET-22**) per model based on 50% subsets of the benchmark items (MAXMIN, MINMIN.*Results over the random samples are averaged over 10 runs. All results are statistically significant ($p < 0.001$).

	cs	de	es	hi	is	ja	ru	uk	zh
All $\leftrightarrow$ MAXMIN	0.99	0.99	0.99	0.96	0.99	0.99	0.99	0.96	0.99
All $\leftrightarrow$ MINMIN	0.98	0.98	0.99	0.98	1.00	0.99	1.00	0.96	0.99
All $\leftrightarrow$ Random*	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99	0.99
<i>st. dev</i>	± 0.005	± 0.004	± 0.004	± 0.005	± 0.0	± 0.005	± 0.003	± 0.006	± 0.005

Table 9: Correlation (Spearman’s ρ) of model rankings, when calculating average performance (**MetricX-24**) per model based on 50% subsets of the benchmark items (MAXMIN, MINMIN, Random). All results are statistically significant ($p < 0.001$).

<i>Metric: COMET-22</i>			
	Δ Std. Dev.	Δ Avg.	Avg. Coord. Distance
MAXMIN	0.26	0.49	0.12
MINMIN	0.18	0.41	0.08
Random*	0.24	0.48	0.11
<i>st. dev</i>	± 0.01	± 0.01	± 0.00

Table 10: Difference between min and max value on the x-axis (Δ Std. Dev.), y-axis (Δ Avg.) and the average coordinate distance for each item in the sampled benchmarks, for WMT24 en $\rightarrow$ de. Results over the random sample are averaged over 10 runs.

<i>Metric: MetricX-24</i>			
	Δ Std. Dev.	Δ Avg.	Avg. Coord. Distance
MAXMIN	9.33	13.26	3.00
MINMIN	8.96	10.94	2.74
Random*	11.15	12.60	3.09
<i>st. dev</i>	± 0.64	± 0.50	± 0.04

Table 11: Difference between min and max value on the x-axis (Δ Std. Dev.), y-axis (Δ Avg.) and the average coordinate distance for each item in the sampled benchmarks, for WMT24 en $\rightarrow$ de. Results over the random sample are averaged over 10 runs.

<i>Target language: Czech</i>			
	Δ Std. Dev.	Δ Avg.	Avg. Coord. Distance
MAXMIN	46.64	96.25	20.71
MINMIN	28.80	54.06	10.17
Random*	43.75	90.32	17.17
<i>st. dev</i>	± 1.72	± 6.08	± 0.36

Table 12: Difference between min and max value on the x-axis (Δ Std. Dev.), y-axis (Δ Avg.) and the average coordinate distance for each item in the sampled benchmarks, for WMT24 en $\rightarrow$ cs. Results over the random sample are averaged over 10 runs.

<i>Spanish</i>			
	Δ Std. Dev.	Δ Avg.	Avg. Coord. Distance
MAXMIN	44.40	94.79	20.95
MINMIN	26.03	56.32	10.07
Random*	41.89	90.11	17.08
<i>st. dev</i>	± 2.39	± 3.93	± 0.43

Table 13: Difference between min and max value on the x-axis (Δ Std. Dev.), y-axis (Δ Avg.) and the average coordinate distance for each item in the sampled benchmarks, for WMT24 en $\rightarrow$ es. Results over the random sample are averaged over 10 runs.

<i>Target language: Hindi</i>			
	Δ Std. Dev.	Δ Avg.	Avg. Coord. Distance
MAXMIN	47.52	100.0	18.78
MINMIN	39.79	43.50	8.46
Random*	47.70	89.68	15.09
<i>st. dev</i>	± 2.2	± 7.81	± 0.41

Table 14: Difference between min and max value on the x-axis (Δ Std. Dev.), y-axis (Δ Avg.) and the average coordinate distance for each item in the sampled benchmarks, for WMT24 en $\rightarrow$ hi. Results over the random sample are averaged over 10 runs.

<i>Target language: Icelandic</i>			
	Δ Std. Dev.	Δ Avg.	Avg. Coord. Distance
MAXMIN	47.97	94.57	18.31
MINMIN	31.63	59.55	8.38
Random*	46.07	90.41	14.90
<i>st. dev</i>	± 1.99	± 4.61	± 0.28

Table 15: Difference between min and max value on the x-axis (Δ Std. Dev.), y-axis (Δ Avg.) and the average coordinate distance for each item in the sampled benchmarks, for WMT24 en $\rightarrow$ is. Results over the random sample are averaged over 10 runs.

<i>Target language: Japanese</i>			
	Δ Std. Dev.	Δ Avg.	Avg. Coord. Distance
MAXMIN	49.73	99.96	20.87
MINMIN	24.71	77.63	10.51
Random*	46.39	95.34	17.68
<i>st. dev</i>	± 2.8	± 4.62	± 0.54

Table 16: Difference between min and max value on the x-axis (Δ Std. Dev.), y-axis (Δ Avg.) and the average coordinate distance for each item in the sampled benchmarks, for WMT24 en $\rightarrow$ ja. Results over the random sample are averaged over 10 runs.

<i>Target language: Russian</i>			
	Δ Std. Dev.	Δ Avg.	Avg. Coord. Distance
MAXMIN	49.24	97.07	21.16
MINMIN	25.64	72.76	11.31
Random*	47.56	91.92	17.90
<i>st. dev</i>	± 1.34	± 5.3	± 0.37

Table 17: Difference between min and max value on the x-axis (Δ Std. Dev.), y-axis (Δ Avg.) and the average coordinate distance for each item in the sampled benchmarks, for WMT24 en $\rightarrow$ ru. Results over the random sample are averaged over 10 runs.

<i>Target language: Chinese</i>			
	Δ Std. Dev.	Δ Avg.	Avg. Coord. Distance
MAXMIN	49.64	99.31	22.46
MINMIN	35.71	64.67	12.44
Random*	48.0	93.14	18.96
<i>st. dev</i>	± 1.15	± 6.07	± 0.4

Table 19: Difference between min and max value on the x-axis (Δ Std. Dev.), y-axis (Δ Avg.) and the average coordinate distance for each item in the sampled benchmarks, for WMT24 en $\rightarrow$ zh. Results over the random sample are averaged over 10 runs.

<i>Target language: Ukrainian</i>			
	Δ Std. Dev.	Δ Avg.	Avg. Coord. Distance
MAXMIN	47.44	95.84	21.86
MINMIN	32.81	57.51	12.30
Random*	46.89	91.90	18.90
<i>st. dev</i>	± 1.3	± 3.75	± 0.36

Table 18: Difference between min and max value on the x-axis (Δ Std. Dev.), y-axis (Δ Avg.) and the average coordinate distance for each item in the sampled benchmarks, for WMT24 en $\rightarrow$ uk. Results over the random sample are averaged over 10 runs.

	cs	de	es	hi	is	ja	ru	uk	zh
Intercept	0.11	0.14	0.12	0.14	0.15	0.09	0.10	0.09	0.10
Lexical Sim.	-0.07	-0.14	-0.11	-0.15	-0.17	-0.11	-0.08	-0.09	-0.07
Syntactic Sim.	-0.02	-0.00	-0.03	-0.02	-0.04	0.01	-0.02	0.00	-0.03
Semantic Sim.	0.01	0.00	0.00	0.00	-0.00	0.01	0.01	0.02	0.01

Table 20: Results for multiple regression on distance matrices for similarity dimensions and average **COMET-22** scores in WMT24, where statistically significant correlations (1,000 permutations) are **highlighted**. Coefficients are rounded to two decimals.

	cs	de	es	hi	is	ja	ru	uk	zh
Intercept	26.56	28.59	25.84	26.32	29.25	21.35	27.42	18.31	16.32
Lex. Sim.	-37.95	-37.15	-27.60	-36.25	-44.16	-33.30	-39.20	-27.17	-17.89
Syn. Sim.	-3.96	-6.65	-5.61	-6.77	-7.20	-4.55	-4.20	-6.22	-5.95
Sem. Sim.	0.58	0.48	1.05	-1.31	-2.12	3.01	-0.65	8.07	5.84

Table 21: Results for multiple regression on distance matrices for similarity dimensions and average **BLEU** scores in WMT24, where statistically significant correlations (1,000 permutations) are **highlighted**. Coefficients are rounded to two decimals.

	cs	de	es	hi	is	ja	ru	uk	zh
Intercept	2.31	3.73	3.17	3.07	5.59	2.78	3.54	2.90	3.02
Lex. Sim.	-1.56	-3.50	-2.43	-3.95	-7.74	-3.53	-2.13	-3.26	-3.66
Syn. Sim.	0.58	0.43	0.51	-0.58	-0.29	-0.62	0.36	0.09	-0.48
Sem. Sim.	0.42	0.44	0.09	0.25	0.445	0.42	0.16	-0.02	0.65

Table 22: Results for multiple regression on distance matrices for similarity dimensions and average **MetricX-24** scores in WMT24, where statistically significant correlations (1,000 permutations) are **highlighted**. Coefficients are rounded to two decimals.

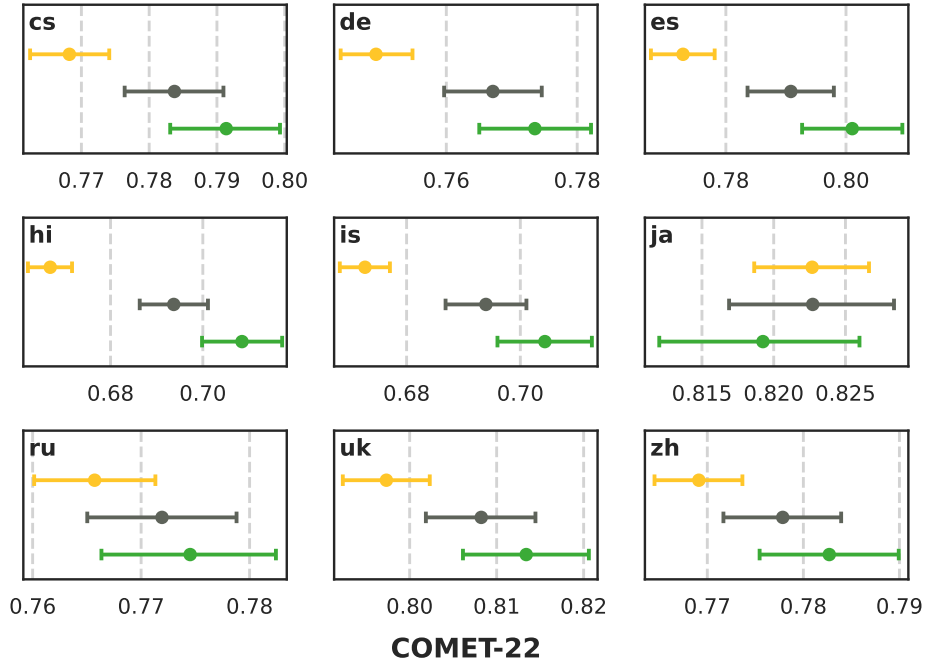


Figure 8: 95% bootstrap confidence intervals (sampled 10,000 times with replacement) for the **MAXMIN** (bottom) and **MINMIN** (top) and Random (middle) for **COMET-22**.

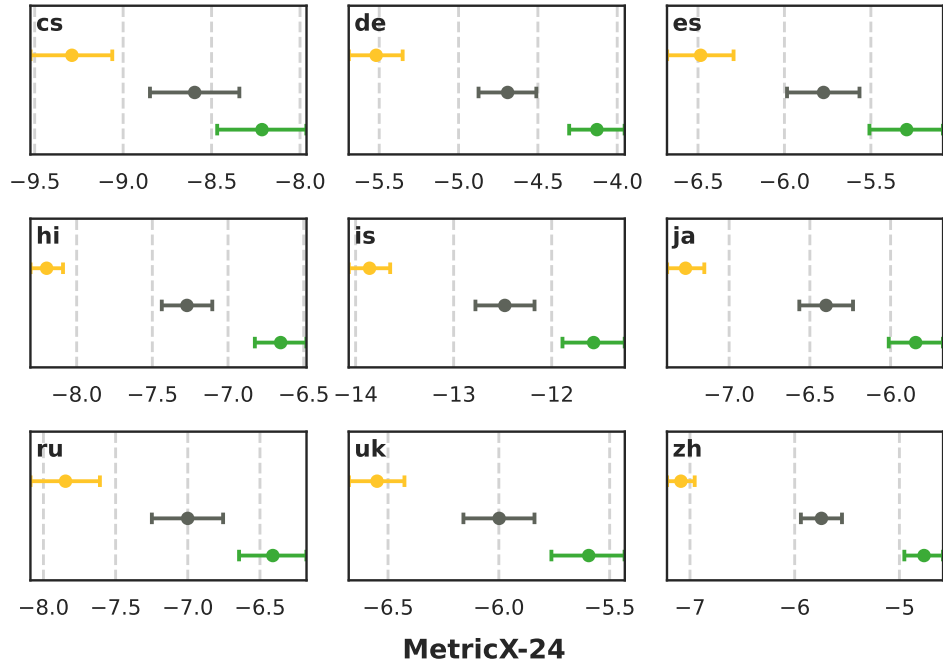


Figure 9: 95% bootstrap confidence intervals (sampled 10,000 times with replacement) for the **MAXMIN** (bottom) and **MINMIN** (top) and Random (middle) for **MetricX-24**.