

AGGREE: An Aggregation-Grounded Challenge Set for Segment-Level MT Metric Evaluation

Lena Griesbeck and Michaela Geierhos

Research Institute CODE, University of the Bundeswehr Munich, Neubiberg, Germany
{lena.griesbeck,michaela.geierhos}@unibw.de

Abstract

This paper introduces AGGREE, a WMT26 Metrics shared task challenge set, which is designed for the diagnostic evaluation of automatic machine translation quality evaluation systems. It is motivated by the observation that metric agreement can be inflated by aggregation as metrics may correlate well at a system level, but disagree at an individual segment level. AGGREE therefore focuses on naturally occurring system outputs where automatic metrics disagree locally or diverge from available WMT human judgments. The challenge set contains 350 single-hypothesis examples across six language pairs drawn from WMT23 and WMT24 system outputs. The submitted set was mined using diagnostics for metric disagreement and metric-human mismatch, then filtered using strict automated quality control. This WMT26 analysis serves only to describe the behavior of the metrics in AGGREE and does not provide an assessment of the performance degradation compared to standard test data.

1 Introduction

Metrics for automatic machine translation (MT) are typically evaluated based on their alignment with human judgments at the segment and system levels. System-level agreement is often substantially higher because scores are averaged over many examples. However, this aggregation can hide local failures. For example, a metric may rank systems well on average while still assigning misleading scores to individual translations. These local failures are important in practice because metrics are also used for model development, data filtering, and fine-grained error analysis.

The AGGREE challenge set is designed to stress-test the reliability of WMT26 Metrics. It targets naturally occurring WMT system outputs where automatic metrics disagree with one another or differ from human judgments of WMT systems. The

set is not designed to estimate the average performance of metrics on the full WMT test distribution. Instead, it focuses on segments where metrics disagree or diverge from human judgments, making it a diagnostic stress test rather than a representative benchmark. The final version submitted contains 350 single-hypothesis examples across six language directions. It uses outputs and human scores from the WMT23 and WMT24 systems from [google-research/mt-metrics-eval](https://github.com/google-research/mt-metrics-eval)¹ and contains 314 MQM-scored rows and 36 ESA-scored rows. Because all three ESA directions contain fewer than 15 examples, the MQM portion forms the empirical core and the ESA rows are treated as supplementary observations. The examples were selected from a larger pool using disagreement diagnostics and filtered with strict automatic text quality checks.

Our contributions are as follows: (1) a disagreement-driven mining pipeline that surfaces naturally occurring regions of local metric disagreement across five diagnostic types rather than using synthetic perturbations, (2) AGGREE, a quality-controlled challenge set with a 314-example MQM core and 36 supplementary ESA examples, and (3) a descriptive analysis of WMT26 Task 2 submissions on this fixed diagnostic sample.

2 Related Work

Traditionally, the validation of machine translation metrics is conducted through meta-evaluation against human judgments, such as direct assessment, multidimensional quality metrics (MQM), or error span annotation (ESA). Annual WMT Metrics and Quality Estimation Shared Tasks benchmark these models using global correlation coefficients (e.g., Pearson’s r , Spearman’s ρ , or Kendall’s τ) at both system and segment levels (Freitag et al.,

¹<https://github.com/google-research/mt-metrics-eval> (last accessed on 2026-08-18)

2022; Zerva et al., 2024). These reports consistently show that system-level correlations are substantially higher than segment-level correlations. However, Mathur et al. (2020) demonstrate that these global correlations are highly sensitive to outlier systems, which can create a false sense of security regarding a metric’s ability to distinguish systems of similar quality. While previous studies have focused on the relationship between metrics and humans, AGGREE adds inter-metric disagreement and metric-human mismatch as diagnostic signals.

To move beyond global scores, several challenge sets have been introduced to test the sensitivity of metrics to specific linguistic phenomena. ACES (Amrhein et al., 2022) uses 68 phenomena based on the MQM ontology and includes synthetic perturbations such as character-level edits, swapping numbers, and swapping antonyms to create minimal pairs. Similarly, DEMETR (Karpinska et al., 2022) uses 35 linguistic categories to demonstrate that, although learned metrics (e.g., COMET and BLEURT) outperform string-based measures, they often fail to distinguish between minor and critical error severities. Unlike these sets, which evaluate metrics against predefined, artificial errors, AGGREE identifies naturally occurring disagreements from actual WMT system outputs where available or submitted metrics disagree locally. Other work targets specific biases or low-resource scenarios. GAMBIT+ (Filandrianos et al., 2025) evaluates gender bias in quality estimation (QE) metrics by providing contrastive masculine and feminine versions of gender-ambiguous occupational terms. SSA-MTE (Li et al., 2025) benchmarks metrics on Sub-Saharan African languages, revealing a significant performance drop in low-resource settings. Zouhar et al. (2024) additionally show that fine-tuned metrics like COMET, despite high WMT correlations, struggle with domain shifts (e.g., to the biomedical domain). These studies confirm that high global performance does not guarantee local or cross-domain reliability. Closely related work, MSLC25 (Knowles et al., 2025), provides an empirical analysis of segment-level inter-metric alignment, observing that automatic measures diverge significantly when evaluating low-quality MT. AGGREE complements these efforts by mining naturally occurring WMT outputs in regions where metrics disagree locally rather than constructing predefined contrastive perturbations.

3 AGGREE

This section describes the design and construction of AGGREE, including its data sources, mining pipeline, and quality-control process.

3.1 Challenge Set Design

AGGREE is based on the idea that high aggregate metric agreement does not ensure reliable segment-level behavior. Aggregation can suppress local deviations when errors cancel out across segments or when many metrics agree on coarse system quality but disagree on individual examples. Therefore, the challenge set mines segments where local metric behavior is informative, such as inter-metric disagreement, disagreement between lexical and neural metric families, number/entity mismatches, and metric-human mismatch candidates. The final set includes five main diagnostic types:

metric false negative The metrics indicate low quality, but the human quality is high.

lexical-high/neural-low Lexical overlap metrics score segments higher than neural or semantic metrics do.

neural-high/lexical-low Neural and semantic metrics score segments higher than lexical overlap metrics do.

number/entity mismatch A mismatch is indicated by simple number or entity-like token diagnostics.

general_high_disagreement Broad metric disagreement remains, but the example does not match a more specific lexical–neural, entity/number, or human-mismatch pattern.

These labels are diagnostic tags used during the construction process, not error annotations that are manually adjudicated. One row can contain more than one phenomenon, while the corrected metadata assigns one primary type for descriptive analysis and preserves additional diagnostic features. The initial mining scheme also included a *metric false positive* category, where metrics assigned high scores to outputs with low human quality. After the post-submission MQM orientation audit, however, no submitted example retained this category as its corrected primary diagnostic type, so it is not reported as a separate type in the final analysis.

3.2 Data Sources and Candidate Mining

Testset	LP	Human score	Source segments
wmt24	en-de	MQM	998
wmt24	en-ja	ESA	998
wmt24	en-zh	ESA	998
wmt24	en-ru	ESA	998
wmt24	ja-zh	MQM	722
wmt23	zh-en	MQM	1,976
wmt23	en-de	MQM	557

Table 1: Selected AGGREE mining settings. Segment counts refer to source segments in the selected MTME setting before challenge-set mining and filtering.

The pipeline uses google-research/mt-metrics-eval as the primary source for WMT system outputs, human evaluations, references, and metric values. Eligible settings were evaluated based on human evaluation and metric coverage, number of systems, and available human-score protocol. WMT24 served as the primary data source because it provides recent system outputs and broad metric coverage. WMT23 was additionally included to ensure sufficient coverage of MQM-based human judgments. Although WMT25 was evaluated, it was not used in the standard core because human evaluation coverage was sparser for this pipeline. Table 1 lists the seven settings used for mining.

The source MQM scores in the four selected MQM settings are non-positive penalties, for which values closer to zero indicate better translation quality, while the selected ESA scores are higher-is-better values between 0 and 100. A post-submission audit found that the initial construction code had negated these already quality-oriented MQM penalties and then used the resulting positive penalty magnitudes as if larger values indicated better quality when computing human-mismatch features. We corrected the direction logic, recomputed all human-dependent features over the complete original (testset, lp) settings, and mapped the corrected metadata back to the unchanged 350 submitted segment identifiers. We used the available metric score columns in google-research/mt-metrics-eval for mining. The selected AGGREE settings included lexical metrics such as BLEU, spBLEU, and chrF; embedding-style metrics such as BERTScore and BLEURT; neural and QE-style metrics such as COMET, COMETKiwi, XCOMET, and MetricX; and several additional WMT metric variants.

Known metric families were normalized into a common higher-is-better direction using conserva-

Language direction		Diagnostic type	
LP	Segs.	Type	Segs.
en-de	120	lexical-high/neural-low	104
zh-en	120	number/entity mismatch	95
ja-zh	74	general high disagreement	89
en-ja	14	neural-high/lexical-low	49
en-ru	11	metric false negative	13
en-zh	11		

Table 2: Final AGGREE distribution after the MQM orientation audit. The corrected private metadata contains 350 segments across six language directions and five primary diagnostic types.

tive name-based heuristics. Metrics whose family or direction could not be inferred were retained for broad disagreement diagnostics but were not used to define the lexical-neural gap. Z-scores and percentile ranks were computed within each (testset, lp) setting, so that score-scale differences across test sets and language pairs (lp) did not dominate candidate selection. Appendix A gives the exhaustive MTME score-field inventory, including version and reference/source suffixes, for every selected setting. Candidate mining computed row-level features for metric-family disagreement, metric-human mismatch, false-positive and false-negative behavior, lexical-neural gaps, length ratios, number mismatches, and entity-like token mismatches. A disagreement-based mining run selected 900 preliminary candidates, 150 from each of six preliminary families, and the finalization stage excluded candidates that carried the pre-audit general-high-disagreement label. Because the pre-audit score affected which hypotheses entered the fixed submission, the corrected metadata does not make the set equivalent to a fully rerun construction pipeline. AGGREE must therefore be interpreted as disagreement-enriched rather than as a balanced sample of correctly oriented metric-human mismatch categories.

3.3 Quality Control and Final Dataset

Challenge-set mining can surface artifacts for the wrong reasons. Very short fragments, URLs, source copies, boilerplate refusals, and near-reference copies may cause disagreement in metrics, but they are usually poor evidence of nuanced metric reliability. Therefore, AGGREE separates discovery from finalization. Broad mining is allowed to surface unusual examples, but the final core must pass stricter text-quality checks. Text-quality filtering checked for reference pseudo-

systems, exact reference copies, source-copy outputs, refusal or non-translation boilerplate, length pathologies, near-identical reference/MT strings, very short fragments, mostly numeric or punctuation strings, low-letter-ratio strings, date or fraction-only strings, placeholder/noise strings, and HTML, UI, URL, or email artifacts. The diagnostic pass inspected 900 candidate rows and marked 243 hard text-quality failures (27.0 %). The most frequent hard-relevant flag was a "too short text" flag, with 202 rows (22.4 %). The final build applied strict hard filtering before soft balancing. This process removed nine source-copy outputs, two refusal or non-translation outputs, 234 remaining text quality failures, 104 rows carrying the pre-audit general-high-disagreement label, and 65 rows capped at a maximum of two rows with the same normalized source. The text quality count is sequential: all nine source-copy outputs were among the 243 hard failures and had already been removed before the remaining 234 were counted. The resulting pool of quality-passing candidates contained 486 entries, of which 350 were selected.

The final distribution is shown in Table 2. Of the test set, 162 rows come from WMT23 and 188 come from WMT24. According to the human-score protocol, 314 rows are MQM-scored, while 36 are ESA-scored. The set is intentionally not uniformly balanced across language pairs because some language pairs only contained a small amount of examples. Specifically, en-de and zh-en contribute 120 rows each; ja-zh contributes 74; and en-ja, en-ru, and en-zh are small diagnostic slices. All 89 examples with corrected primary type "general high disagreement" had pre-audit metric-human mismatch labels: 53 were labeled metric false negatives and 36 metric false positives. After correction, no other diagnostic tag applies to these rows.

AGGREE is a single-hypothesis approach rather than a contrastive one. A paired-hypothesis feasibility analysis was conducted to test if each example could be paired with a second output that passed QC from another system for the same source segment. The original margin of 0.30 on the within-setting human-quality percentile scale was chosen as a conservative operational threshold to avoid treating near-tied human scores as a meaningful contrast. Even at a permissive margin of 0.15, only 14 of 350 examples (4.0%) were pairable, with no language direction reaching ten pairs (7 at 0.30, 9 at 0.20).

4 Challenge-Set Results on WMT26 Metric Submissions

We analyze the final WMT26 metric outputs returned for AGGREE as a diagnostic evaluation, not an official WMT ranking. The results of the challenge set contain Task 2 quality score predictions. All predictions were linked by segment identifier and language pair to the corrected AGGREE metadata. Since no suitable random data sets were evaluated from the WMT26 submissions, these correlations describe only the relationship with human quality assessment on AGGREE. They do not measure a relative decline in performance compared to standard WMT data. The coefficients correlate the submitted metric values with the rankings from the human quality assessment. However, AGGREE itself was selected in part based on historical discrepancies between metrics and human evaluation, as well as on mismatches between metrics and human evaluation. Therefore, the results should be interpreted as behavior observed in a deliberately conditioned diagnostic sample.

For each submitted score, we compare the metric output of the hypothesis to a hidden human-quality percentile rank. The human target is computed once within each (testset, lp) slice of the final 350-example set, independently of submission coverage, with tied values assigned their average rank. The source MQM values are non-positive penalties and are retained because values closer to zero indicate better quality, while ESA values are retained in their original higher-is-better direction. Thus, larger human-quality percentile ranks always indicate higher translation quality ratings. Because raw submitted metric scales may differ systematically across language pairs, each submission's scores are converted to within-(testset, lp) percentile ranks before computing the pooled correlation. This within-setting Spearman correlation is used as the primary reported statistic, and the raw-score Spearman is retained as a robustness column.

Table 3 shows that the strongest full-coverage association after within-setting score normalization is obtained by Vertical_870257 ($\rho_{\text{within}} = 0.787$), followed by Vertical_870554 (0.744), FACET_869546 (0.602), and FACET_869543 (0.601). The same systems remain among the strongest under the raw-score Spearman analysis, which indicates that the main pattern is robust to this normalization check. However, the order of the top two systems switches. Restricting

Submission	LPs	ρ_{within}	ρ_{raw}
Vertical_870257	6	0.787	0.678
Vertical_870554	6	0.744	0.689
FACET_869546	6	0.602	0.613
FACET_869543	6	0.601	0.610
xCOMET XL Reference-free - Baseline	4	0.571	0.603
Gemini-3.6-Flash	4	0.556	0.509
Cohere CAT+ ensemble	6	0.530	0.546
AEGIS	6	0.524	0.587
xCOMET XXL Reference-free - Baseline	4	0.519	0.553
Gemma 4 Reference-based - Reasoning Baseline	4	0.517	0.455
Gemma 4 - Reasoning Baseline	4	0.509	0.468
Qwen 3.6 - Reasoning Baseline	4	0.504	0.483
fluency2-gemini35-esa	6	0.496	0.518
xCOMET XL Reference-based - Baseline	4	0.495	0.579
MQM-LLM Confidence Aware	6	0.473	0.517
bytepop_pro10	6	0.469	0.470
xCOMET XXL Reference-based - Baseline	4	0.464	0.530
bytepop_pro9	6	0.462	0.467
Qwen 3.6 Reference-based - Reasoning Baseline	4	0.459	0.436
MQM-LLM	6	0.449	0.496
gemba-poly	6	0.439	0.437
COMETKiwi22	6	0.412	0.376
COMET22	6	0.359	0.356
YiSi-1	6	0.238	0.185
BERTScore	6	0.232	0.098
spBLEU	6	0.136	0.076
BLEU	6	0.130	0.100
chrF	6	0.060	0.012
cuni-v14-regression	1	0.050	0.050

Table 3: Task 2 Spearman correlations over all returned submissions with at least one AGGREE score. ρ_{within} first converts each submission’s scores to percentile ranks within each (testset, lp) slice before pooling. ρ_{raw} uses the raw submitted scores. Both are correlated with quality-oriented human percentile ranks from the corrected metadata. Correlations based on different language-pair subsets are descriptive and are not directly comparable as an overall metric ranking.

the analysis to the 314 MQM rows preserves this ordering among submissions that cover all three sufficiently sized MQM directions (en-de, zh-en, and ja-zh): the corresponding within-setting Spearman correlations are 0.819, 0.760, 0.690, and 0.689. We do not report correlations for the 36 ESA rows alone because their three language directions contain only 14, 11, and 11 examples. We use the Spearman correlation because our analysis focuses on the relative ordering of segments based on human ratings. Within-setting normalization prevents differences in submitted score distributions across languages from determining the pooled ordering. Several strong submissions cover only the four English-source directions. Their correlations are therefore based on different subsets from those of the six-direction submissions and should not be interpreted as directly comparable overall rankings.

Table 4 summarizes the distribution of metric behavior across diagnostic types. This is closer to the central AGGREE question: whether submitted met-

rics behave differently across particular disagreement mechanisms. The median association is highest for number/entity mismatches (0.384), followed by neural-high/lexical-low (0.342) and general-high-disagreement cases (0.312), whereas lexical-high/neural-low examples have a near-zero median correlation (0.024). The corrected metric-false-negative subset contains only 13 examples and is therefore excluded from the correlation-distribution summary. These descriptive differences support type-conditioned analysis, but they do not establish general metric performance because both the original selection procedure and the post-audit labels are construction-dependent. We do not use per-language-pair maxima to claim language-pair effects. The maxima can be produced by different submissions, the English-to-Japanese, English-to-Russian, and English-to-Chinese slices contain only 14, 11, and 11 examples, and the current design does not separate submission effects from language-pair effects. Language-pair-specific values are therefore exploratory diagnostics only.

Diagnostic type	Examples	Median ρ	IQR
lexical-high/neural-low	104	0.024	[-0.136, 0.159]
number/entity mismatch	95	0.384	[0.224, 0.610]
general high disagreement	89	0.312	[0.285, 0.350]
neural-high/lexical-low	49	0.342	[0.187, 0.415]
metric false negative	13	–	–

Table 4: Distribution of within-setting-normalized Task 2 Spearman correlations by corrected AGGREE diagnostic type. Median and interquartile range summarize submissions with at least 20 scored examples for the type. The 13-example metric-false-negative subset is shown for completeness but is not summarized.

5 Discussion

The returned WMT26 outputs provide a descriptive view of metric behavior on AGGREE. Coverage differs across submissions, so correlations based on different language subsets are not directly comparable. Among submissions with the same full coverage, associations with human quality differ, and the diagnostic-type breakdown identifies where these differences are concentrated. AGGREE should therefore be used alongside broad representative evaluations, with its results interpreted as behavior on a deliberately selection-conditioned sample.

6 Conclusion

AGGREE is an aggregation-grounded WMT26 challenge set containing 350 quality-filtered, single-hypothesis examples derived from human-rated WMT23/24 outputs. Its main empirical basis is a 314-example MQM portion, 36 ESA examples are retained as small supplementary slices. The returned WMT26 results describe heterogeneous behavior on this fixed diagnostic sample. These results do not constitute an independent general-purpose metric ranking or a comparison with normal-data performance.

7 Limitations

AGGREE is automatically quality controlled and grounded in original WMT human scores, but it is not independently and exhaustively adjudicated by human annotators. Its diagnostic types are construction labels rather than gold error categories. The small en-ja, en-ru, and en-zh slices contain all 36 ESA examples, so the substantive analysis is primarily MQM-based.

The post-submission audit corrected MQM-dependent metadata and changed 98 primary diagnostic labels, but it cannot undo the influence of the pre-fix human-mismatch features on which hypotheses entered the fixed submission. Moreover,

human judgments and historical metric scores were used in selection and evaluation, and several returned submissions overlap by name or family with construction metrics. The resulting correlations should therefore be read as diagnostic behavior on an intentionally mined stress test, not as general segment-level unreliability.

No matched random set was scored by the returned WMT26 submissions, and a new common scoring round was not available after the shared-task scoring period. The pairing sensitivity analysis is also limited to the inspected 900-candidate review pool, so it supports the submitted single-hypothesis design within that pool but does not rule out a larger paired set from a broader future search.

Future work should add an organizer-scored matched control set, independent manual adjudication, and a larger contrastive set if sufficient paired examples can be obtained.

References

- Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. [ACES: Translation accuracy challenge sets for evaluating machine translation metrics](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- George Filandrianos, Orfeas Menis Mastromichalakis, Wafaa Mohammed, Giuseppe Attanasio, and Chrysoula Zerva. 2025. [GAMBIT+: A challenge set for evaluating gender bias in machine translation quality estimation metrics](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 314–326, Suzhou, China. Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. [Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust](#). In *Proceedings of the Seventh Conference*

on Machine Translation (WMT), pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. [DEMETR: Diagnosing evaluation metrics for translation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rebecca Knowles, Samuel Larkin, and Chi-Kiu Lo. 2025. [MSLC25: Metric performance on low-quality machine translation, empty strings, and language variants](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 945–956, Suzhou, China. Association for Computational Linguistics.

Senyu Li, Felermario Dario Mario Ali, Jiayi Wang, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenertorp, Colin Cherry, and David Ifeoluwa Adelani. 2025. [Evaluating WMT 2025 metrics shared task submissions on the SSA-MTE African challenge set](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 913–919, Suzhou, China. Association for Computational Linguistics.

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

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. [Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE?](#) In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. [Fine-tuned machine translation metrics struggle in unseen domains](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500, Bangkok, Thailand. Association for Computational Linguistics.

A Mining Metric Inventory

This appendix lists the complete MTME metric-score inventory available to the AGGREE mining pipeline after filtering to human-scored outputs in the seven selected settings. Names are reproduced exactly from the local MTME exports.

The WMT26 evaluation correlations reported in Section 4 are not correlations against these mining metrics. They correlate submitted WMT26 scores with corrected human-quality ranks. However, because the challenge set was selected partly using the historical scores listed here, AGGREE remains a selection-conditioned diagnostic sample rather than an independent representative benchmark.

WMT23 en-de and zh-en. Both WMT23 settings contain the same 68 metric fields.

Reference-based refA fields (25)

- BERTscore-refA
- BLEU-refA
- BLEURT-20-refA
- COMET-refA
- Calibri-COMET22-refA
- MEE4-refA
- MaTESe-refA
- MetricX-23-b-refA
- MetricX-23-c-refA
- MetricX-23-refA
- XCOMET-Ensemble-refA
- XCOMET-XL-refA
- XCOMET-XXL-refA
- XLsim-refA
- YiSi-1-refA
- chrF-refA
- docWMT22CometDA-refA
- eBLEU-refA
- embed_llama-refA
- f200spBLEU-refA
- instructscore-refA
- mre-score-labse-regular-refA
- prismRef-refA
- sescoreX-refA
- tokengram_F-refA

Synthetic-reference fields (25)

- BERTscore-synthetic_ref
- BLEU-synthetic_ref
- BLEURT-20-synthetic_ref
- COMET-synthetic_ref
- Calibri-COMET22-synthetic_ref
- MEE4-synthetic_ref
- MaTESe-synthetic_ref
- MetricX-23-b-synthetic_ref
- MetricX-23-c-synthetic_ref
- MetricX-23-synthetic_ref
- XCOMET-Ensemble-synthetic_ref
- XCOMET-XL-synthetic_ref
- XCOMET-XXL-synthetic_ref
- XLsim-synthetic_ref
- YiSi-1-synthetic_ref
- chrF-synthetic_ref
- docWMT22CometDA-synthetic_ref
- eBLEU-synthetic_ref
- embed_llama-synthetic_ref
- f200spBLEU-synthetic_ref
- instructscore-synthetic_ref
- mre-score-labse-regular-synthetic_ref
- prismRef-synthetic_ref
- sescoreX-synthetic_ref
- tokengram_F-synthetic_ref

Source-based/QE-style fields (18)

- Calibri-COMET22-QE-src
- CometKiwi-XL-src
- CometKiwi-XXL-src
- CometKiwi-src
- GEMBA-MQM-src
- KG-BERTScore-src
- MS-COMET-QE-22-src
- MetricX-23-QE-b-src
- MetricX-23-QE-c-src
- MetricX-23-QE-src
- Random-sysname-src
- XCOMET-QE-Ensemble-src
- cometoid22-wmt21-src
- cometoid22-wmt22-src
- cometoid22-wmt23-src
- docWMT22CometKiwiDA-src
- mbr-metricx-qe-src
- prismSrc-src

WMT24 shared fields. The WMT24 en-de, en-ja, en-ru, en-zh, and ja-zh settings all contain the following 31 shared fields.

Reference-based refA fields (19)

- BERTScore-refA
- BLEU-refA
- BLEURT-20-refA
- COMET-22-refA
- MetricX-24-Hybrid-refA
- MetricX-24-refA
- PrismRefMedium-refA
- PrismRefSmall-refA
- XCOMET-refA
- YiSi-1-refA
- chrF-refA
- chrFS-refA
- damonmonli-refA
- metametrics_mt_mqm_hybrid_kendall-refA
- metametrics_mt_mqm_kendall-refA
- metametrics_mt_mqm_same_source_targ-refA
- monmonli-refA
- sentinel-ref-mqm-refA
- spBLEU-refA

Source-based/QE-style fields (12)

- CometKiwi-XXL-src
- CometKiwi-src
- MetricX-24-Hybrid-QE-src
- MetricX-24-QE-src
- XCOMET-QE-src
- XLsimDA-src
- XLsimMqm-src
- gemba_esa-src
- metametrics_mt_mqm_qe_kendall_seg_s-src
- metametrics_mt_mqm_qe_same_source_t-src
- sentinel-cand-mqm-src
- sentinel-src-mqm-src

Additional WMT24 ja-zh and en-de fields. The WMT24 ja-zh and en-de settings contain four additional fields. Together with the 31 shared WMT24 fields, these yield 35 fields for ja-zh.

- BLCOM-refA
- BLCOM_1-refA
- MEE4-refA
- bright-qe-src

Additional WMT24 en-de-only fields. The WMT24 en-de setting additionally contains 22 refB fields. Together with the 31 shared fields and the four additional fields above, these yield 57 fields for WMT24 en-de.

- BERTScore-refB
- BLCOM-refB
- BLCOM_1-refB
- BLEU-refB
- BLEURT-20-refB
- COMET-22-refB
- MEE4-refB
- MetricX-24-Hybrid-refB
- MetricX-24-refB
- PrismRefMedium-refB
- PrismRefSmall-refB
- XCOMET-refB
- YiSi-1-refB
- chrF-refB
- chrFS-refB
- damonmonli-refB
- metametrics_mt_mqm_hybrid_kendall-refB
- metametrics_mt_mqm_kendall-refB
- metametrics_mt_mqm_same_source_targ-refB
- monmonli-refB
- sentinel-ref-mqm-refB
- spBLEU-refB