

AmanaMT: Translation Direction Predicts Automatic Metric Reliability

Isaac Thompson
Brigham Young University
it238@byu.edu

Abstract

When machine translation (MT) metrics fail in low-resource settings, the field typically blames data scarcity. We show this diagnosis is often wrong using AmanaMT, a unified benchmark spanning 730,748 segments, 33 languages, three annotation tiers, and 10 metrics: the dominant predictor of metric unreliability is not resource level but *domain mismatch* and *translation direction*. Direction is the single largest predictor of between-language variance in metric reliability ($\eta_{\text{dir}}^2 = 0.309$, nearly a third of between-language variance); morphological type is a substantially smaller main effect ($\eta_{\text{morph}}^2 = 0.073$), though the joint direction $\times$ morphology cells account for $\eta_{\text{cell}}^2 = 0.618$, capturing direction-specific morphological variation. Ukrainian’s near-zero COMET correlation is fully explained by a domain label (other, not news): a genre confound, not a language problem. Russian’s weak aggregate dissolves into clean per-year signals once stratified by WMT year, a textbook Simpson’s Paradox from pooling heterogeneous system generations. Neural metrics (MetricX-24, xCOMET, COMET) substantially outperform n -gram baselines at every tier; the gap is largest in morphologically complex and non-news settings where surface overlap is an especially poor proxy for adequacy.

1 Introduction

Automated MT metrics are the connective tissue of modern MT research: they drive model selection, benchmark comparisons, and shared-task rankings across hundreds of language pairs. Yet the conditions under which these metrics are trustworthy remain poorly understood. A practitioner evaluating a new Hindi or Swahili translation system reaches for BLEU or COMET without knowing whether those metrics bear any reliable relationship to human judgment for that language, domain, or translation direction.

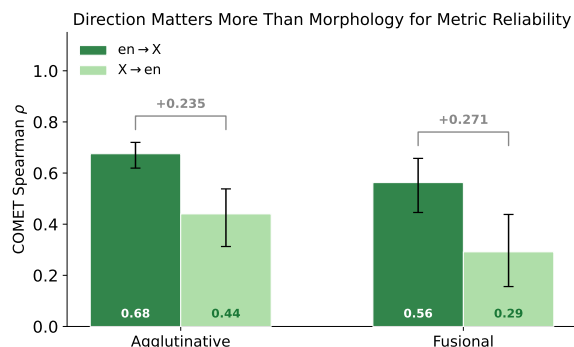


Figure 1: Translation direction matters more than morphological type for COMET reliability: en $\rightarrow$ X (dark green) consistently outperforms X $\rightarrow$ en (light green) within both morphological types with Tier-1 data in both directions (§3.3). Isolating is omitted: there are no Tier-1 en $\rightarrow$ X isolating observations (Limitations), so it cannot show a direction contrast.

The WMT Metrics Shared Task (Freitag et al., 2023) has produced annual meta-evaluations for over a decade, but each edition covers only the language pairs in that year’s shared task, typically 5–15 pairs, nearly always news-domain, and never with a systematic typological framing. Efforts to extend coverage to low-resource languages (Yari et al., 2025) remain limited in scope. No prior work has unified professional Multidimensional Quality Metrics (MQM) annotations, crowd-sourced Direct Assessment (DA), and synthetic annotations across a common evaluation framework spanning the full range of typological diversity.

We fill this gap with AmanaMT¹, a unified three-tier benchmark across 33 languages, 10 metrics, and 730,748 segments. Our design enables the first direct comparison of metric reliability across resource tiers, translation directions, and morpholog-

¹Amana (Hausa, borrowed from Arabic *amāna*): trust, reliability, honesty. Named for the benchmark’s purpose: determining when and where automatic MT metrics can be trusted. Code and data: <https://github.com/byu-matrix-lab/amana-mt>.

ical types in a single controlled study. The results reveal that the field’s go-to explanation for metric failure, “this language is low-resource,” is often wrong in a correctable way: domain mismatch and the joint configuration of translation direction and morphological type predict failure far better than resource level alone.

1.1 Contributions

Direction dominates over morphological type.

We quantify metric reliability across a 2×3 grid of translation direction ($\text{en} \rightarrow \text{X}$, $\text{X} \rightarrow \text{en}$, $\text{X} \leftrightarrow \text{Y}$) and morphological type (isolating, agglutinative, fusional). Translation direction is the dominant predictor ($\eta_{\text{dir}}^2 = 0.309$, nearly a third of between-language variance); morphological type is a much smaller main effect ($\eta_{\text{morph}}^2 = 0.073$). The joint direction $\times$ morphology cells account for $\eta_{\text{cell}}^2 = 0.618$, capturing direction-specific morphological variation (particularly within $\text{X} \rightarrow \text{en}$) absent from prior MT meta-evaluation work.

Domain as the primary confound. Ukrainian’s near-zero correlation ($\rho = .001$) is fully explained by a domain label mismatch (`uk-en_2022_other`): every Ukrainian segment is conversational-register text evaluated against a news-trained metric. Controlled comparison with direction-matched $\text{X} \rightarrow \text{en}$ languages confirms a 0.143 absolute gap to the next-lowest language (`ja-en`), attributable to domain shift, not language difficulty.

Simpson’s Paradox in multi-year evaluation. Russian’s anomalously low aggregate correlation ($\rho = .157$) conceals clean per-year signals ($\rho = .383$ in 2021, $.501$ in 2022). Pooling WMT system generations across years creates aggregation bias that should be treated as a methodological artefact, not a linguistic finding.

Neural metrics dominate across typological cells and resource tiers. MetricX-24, xCOMET, and COMET outperform surface metrics in every typological cell and at every resource tier. The gap is largest for the hardest cases: morphologically complex languages and non-news domains.

2 Related Work

WMT Metrics Shared Task. The primary venue for systematic MT metric evaluation is the annual WMT Metrics Shared Task, which has evolved from BLEU-focused comparisons (Bojar et al., 2017) to MQM-based meta-evaluation (Freitag et al., 2023). Each edition covers 5–15 language

pairs with professional MQM annotations, generating yearly rankings. Our work differs in three ways: (1) we aggregate across years rather than evaluating a single shared-task snapshot; (2) we include 22 additional language pairs via DA and synthetic annotations; (3) we frame results typologically rather than as a ranking exercise.

MQM and annotation quality. Freitag et al. (2021) establish that expert MQM annotation substantially outperforms crowd-sourced DA for high-quality MT evaluation. Deutsch et al. (2023) show that tie handling in pairwise accuracy (PA) materially affects rankings, motivating our adoption of soft pairwise accuracy (SPA) (Thompson et al., 2024). Reference quality as a confound is studied by Freitag et al. (2023), who argue that “metrics might be guilty but references are not innocent.”

Metric reliability under domain and resource shift. Zouhar et al. (2024) demonstrate that learned metrics fine-tuned on WMT news data degrade substantially when applied to out-of-domain text, directly predicting the Ukrainian result we observe. Yari et al. (2025) find that metric reliability varies significantly across Indian languages, with low-resource languages showing systematically weaker correlations, a finding our broader typological framing partially confirms and partially complicates (Khmer is low-resource but correlates well due to domain homogeneity).

Morphology and MT evaluation. Mondshine et al. (2025) study metric reliability across morphological types in summarization and find that COMET correlates better for fusional languages, the inverse of the sign we observe within $\text{X} \rightarrow \text{en}$. In MT, morphological type is not an independent main driver of metric reliability ($\eta_{\text{morph}}^2 = 0.073$); the direction axis, absent from monolingual summarization, dominates instead, and morphology’s residual effect only surfaces as a within-direction interaction rather than a standalone predictor. Bugliarello et al. (2020) show that MT difficulty (cross-mutual information) is asymmetric across translation direction; we show the same asymmetry holds for metric reliability, extending their finding from difficulty to evaluability.

Typological diversity in NLP. The broader context for this work is the recognition that NLP systems, including evaluation metrics, are not language-agnostic (NLLB Team and Costa-jussà, 2022). The morphological and directional effects

we quantify are one instance of a general pattern: models trained predominantly on English-centric data inherit English-centric assumptions about what good translation looks like.

3 Methodology

3.1 Three-Tier Annotation Design

A core contribution is the unified three-tier data pipeline that brings together human and synthetic annotations under a shared evaluation framework:

Tier 1a, WMT MQM (*high-resource*): Professional Multidimensional Quality Metrics span-level annotations from the WMT shared task (Freitag et al., 2021), covering de, zh, ru, he (WMT20–23). Segment-level MQM penalty scores are converted to sentence-level quality scores via the standard weighted formula. WMT24 en-es also uses MQM annotation but is tiered as Medium, not High (see §4.2), since it is single-year with a more recent system generation and annotation cohort than de/zh/ru/he.

Tier 1b, WMT Direct Assessment (*medium/low-resource*): Sentence-level DA scores from the WMT direct assessment protocol (Graham et al., 2017) covering cs, tr, uk (medium) and ha, km, ps (low-resource), drawn from the corresponding per-year WMT shared task releases. DA scores are normalized to [0, 1].

Tier 2, Synthetic MQM (*very-low-resource*): For 7 languages with no human MT annotations (sw, ht, lo, th, my, am, ka), we generate hypotheses using NLLB-200 (NLLB Team and Costa-jussà, 2022) over FLORES+ (Goyal et al., 2022; NLLB Team and Costa-jussà, 2022) source sentences (openlanguagedata/flores_plus, the community-maintained successor to FLORES-101 that also incorporates the FLORES-200 expansion introduced alongside NLLB-200), then prompt the large language model (LLM) Llama-3.1-8B-Instruct (Dubey et al., 2024) with a GEMBA-MQM-style few-shot template (Kocmi and Federmann, 2023a) to produce synthetic span-level quality judgments. **Circular-evaluation caveat:** the same Llama-3.1-8B backbone is also evaluated as the GEMBA metric. GEMBA correlations are therefore excluded from all Tier 2 analyses; only reference-based and kiwi metrics are reported for this tier. Validation against human annotation (even 200 Error Span Annotation (ESA)-annotated segments per language) would be required to confirm that Tier 2 synthetic scores reflect real translation

quality rather than LLM self-consistency.

3.2 Metrics Evaluated

Lexical surface metrics. **BLEU** (Papineni et al., 2002): n -gram precision with brevity penalty. **ChrF++** (Popović, 2015): Character n -gram F-score with added word-unigram/bigram terms (sacrebleu word_order=2, the chrF++ configuration); more morphology-sensitive than BLEU.

Embedding-based. **BERTScore** (Zhang et al., 2020): Contextual embedding cosine similarity (mDeBERTa-v3-base). **BLEURT** (Sellam et al., 2020): Fine-tuned BERT regression on human ratings.

Learned regression (reference-based). **COMET** (Rei et al., 2022a): Regression over source+hypothesis+reference (wmt22-comet-da). **xCOMET** (Guerreiro et al., 2024): COMET with MQM error-span detection (XCOMET-XL). **MetricX-24** (Juraska et al., 2024): Google’s WMT 2024 winning hybrid metric (metricx-24-hybrid-large-v2p6-bfloat16).

Reference-free quality estimation. **COMET-Kiwi** (Rei et al., 2022b) (Kiwi-22): Reference-free QE using source+hypothesis only (wmt22-cometkiwi-da). **COMET-Kiwi-23** (Rei et al., 2023) (CK23): Larger 2023 variant (wmt23-cometkiwi-da-xl).

LLM-based. **GEMBA-MQM** (Kocmi and Federmann, 2023b,a): Llama-3.1-8B-Instruct (Dubey et al., 2024) prompted with a GEMBA-MQM few-shot template to predict MQM error spans; excluded from Tier 2 analysis (see §3.1).

Ensemble: Min-max normalized score average over the five neural metrics (COMET, xCOMET, MetricX-24, Kiwi-22, Kiwi-23), excluded from Tier 2.

3.3 Translation Direction and Morphological Type Analysis

A central novel contribution is a 2×3 cell analysis of *translation direction* crossed with *morphological type*, defined at the language level as follows.

Translation direction:

X→**en** non-English source, English target (e.g. zh, he, uk)
en→**X** English source, non-English target (e.g. de, tr, sw)

$X \leftrightarrow Y$ neither source nor target is English; each pair is labeled by its target language (e.g. fr for de $\rightarrow$ fr, bn for hi $\rightarrow$ bn, hi for bn $\rightarrow$ hi)

Morphological type (standard linguistic taxonomy):

isolating minimal inflection, meaning via word order (zh, th, lo, km, my)

agglutinative transparent suffix/prefix stacking (tr, fi, et, ka, kk, sw, ja, ta, am)

fusional multiple grammatical features fused per morpheme (de, ru, es, cs, fr, pl, uk, he, ps, bn, hi, gu, and others)

Khmer (km) is Austroasiatic/Mon-Khmer and isolating, not agglutinative as sometimes assumed; Burmese (my) is predominantly analytic. Hausa (ha) is classified fusional here: although it has real agglutinative affixation, its Afroasiatic verbal-grade system fuses aspect, transitivity, and direction via tone and vowel apophony rather than discrete morphemes, and its typology is genuinely mixed rather than a clean fit to either category.

For each language we compute Spearman ρ between each metric and human scores, yielding a 2×3 matrix of mean ρ per (direction, morphology) cell. Effect sizes are reported as η^2 (proportion of variance explained), with 95% bootstrap CIs over the languages in each cell. η_{dir}^2 and η_{morph}^2 are computed via one-way ANOVA on the per-language ρ values with direction and morphological type as the respective factors. η_{cell}^2 is computed via one-way ANOVA over the six direction $\times$ morphology cells and therefore measures total between-cell variance rather than a formally partitioned interaction term; it is bounded below by each main effect by construction. The informative comparison is therefore how direction dominates the main-effect decomposition (η_{dir}^2 much larger than η_{morph}^2), while the cell structure (η_{cell}^2) captures additional direction-specific morphological variation not visible in either main effect alone.

This analysis is motivated by and extends [Mondshine et al. \(2025\)](#), who study metric reliability across morphological types in summarization (not MT) and find that COMET correlates *better* for fusional languages in their setting, treating morphology as an independent main driver. Our result does not simply reverse their sign under an added direction confound: morphology is a substantially weaker independent main driver of metric reliability in MT than direction is ($\eta_{\text{morph}}^2 = 0.073$, versus

$\eta_{\text{dir}}^2 = 0.309$ for direction). Their morphology-driven story does not replicate as a standalone effect in our MT setting; what residual morphological signal exists is direction-dependent (§3.3), a structurally different claim from theirs. The direction axis is independently motivated by [Bugliarello et al. \(2020\)](#), who show that translation *difficulty* (measured via cross-mutual information) is asymmetric across direction; our contribution is showing the same asymmetry holds for metric *reliability*.

3.4 Evaluation Protocol

For each (language, metric) pair, we compute: (1) Spearman ρ with 95% bootstrap confidence intervals (1000 resamples, percentile method); (2) Kendall τ ; (3) Pairwise accuracy (PA), fraction of concordant segment pairs, excluding metric-tied and human-tied pairs ([Freitag et al., 2023](#)); (4) Soft pairwise accuracy (SPA), the WMT 2024 standard ([Thompson et al., 2024](#)), which credits metric-tied pairs as 0.5 rather than excluding them, providing a tie-calibrated alternative to PA ([Deutsch et al., 2023](#)); (5) Williams (1959) test for significance of the difference between two correlations sharing a common criterion.

Segments where the human quality score is tied are excluded from both PA and SPA denominators, following [Deutsch et al. \(2023\)](#).

All main analysis runs over 723,664 professionally annotated segments across Tier 1a and Tier 1b; the 730,748 total segment count includes Tier 2 synthetic annotations, which are reported separately and flagged as exploratory.

4 Results and Analysis

4.1 Main Correlation Results

Table 1 reports per-tier Spearman ρ and soft pairwise accuracy (SPA) averaged across all languages in each tier. Neural regression metrics dominate at every resource level; surface metrics degrade more sharply on Tier 1a (MQM) where human judges penalize subtle errors that n -gram overlap misses. Williams tests (COMET vs. BLEU and vs. chrF, per language) confirm COMET’s advantage over surface metrics is statistically significant at every Tier 1 language (Tier 1a and Tier 1b) except uk, zu: Ukrainian is the domain-confound outlier discussed below, where COMET’s own correlation collapses toward zero so it cannot significantly exceed anything; Zulu’s COMET advantage is significant against BLEU ($p = 7.5 \times 10^{-9}$) but not

against chrF ($p = 3.6 \times 10^{-1}$).

Tier 2 synthetic correlations are near-zero across all metrics, most consistent with the local LLM judge (Llama-3.1-8B) failing to produce reliable quality judgments for these low-resource, largely non-Latin-script languages, rather than with the MT metrics themselves failing. This uniform near-zero pattern does not itself distinguish which metric is more reliable for these languages, and we draw no such conclusion from it; without human validation of the synthetic judgments, we cannot separate this LLM-judge-unreliability explanation from a genuine ceiling on metric reliability at this resource level.

We explicitly ruled out a data-pipeline explanation: Tier 2’s segment identifiers are shared across all seven languages (each keyed only by FLORES+ sentence index, with no per-language disambiguator), a structurally identical risk to the reference/segment-ID pipeline error affecting the he-en and en-es span-fallback pairs, discussed below (§4.2) and in Limitations. Independently rebuilding each language’s quality scores from the raw synthetic-annotation cache and rejoining by language and segment confirms the reported Tier 2 correlations are correctly aligned, not an artifact of cross-language score bleed. We report Tier 2 for completeness, not as evidence for any typological claim.

4.2 Per-Language Analysis

Figure 2 shows per-language Spearman ρ for COMET. Notable patterns:

Top performers: Latvian achieves the highest individual COMET correlation ($\rho = .746$), followed by Lithuanian (0.742), Finnish (0.739), Turkish (0.724), and Estonian (0.721) – all en→X-specific values for the languages with data in both directions (Table 8), not the direction-pooled averages, which would understate them – spanning two morphological types. The unifying feature is not morphological type but data quality: all five use single-year WMT DA news data with high inter-annotator agreement, suppressing the multi-year pooling variance that depresses Russian and German.

Ukrainian collapses to near-zero correlation for all metrics ($\rho_{\text{COMET}} = .001$, pairwise acc. = 0.499). A uniform near-zero result across unrelated metric families (surface overlap and learned regression alike) is also the signature of a broken human-score join, so before accepting a domain

Family	Metric	High	Med	Low
<i>Spearman ρ / SPA</i>				
Lexical	BLEU	.186 / .582	.293 / .636	.230 / .576
	ChrF	.201 / .588	.295 / .632	.267 / .590
Embed.	BERTScore	.265 / .617	.404 / .688	.283 / .595
	BLEURT	.349 / .654	.423 / .702	.291 / .600
Regr.	COMET	.436 / .690	.526 / .747	.331 / .613
	xCOMET	.485 / .709	.496 / .730	.315 / .607
	MetricX-24	.465 / .698	.588 / .773	.383 / .632
QE	Kiwi-22	.380 / .661	.445 / .701	.289 / .598
	Kiwi-23	.380 / .660	.474 / .714	.368 / .626
LLM	GEMBA	.035 / .501	–	–
–	Ensemble	.485 / .709	.557 / .759	.363 / .624

Table 1: Average Spearman ρ / soft pairwise accuracy (SPA) by resource tier. High = Tier 1a MQM (de, ru, zh, he); Med = Tier 1b DA plus en-es WMT24 MQM; Low = Tier 1b DA. Values are averaged across annotation tiers within each resource tier. **Bold:** best single metric per tier. GEMBA shown for MQM tier only; excluded from Tier 1b (Med, Low) because its backbone was not fine-tuned on DA-style human judgments, and from Tier 2 to avoid circular evaluation with the Llama-3.1-8B-generated hypotheses. GEMBA’s High-tier value averages only de/ru/zh, not all four Tier 1a languages: he-en’s GEMBA score predates a segment-identifier correction applied to every other metric (§4.2) and could not be realigned, so it is excluded pending recomputation (see Limitations). en-es’s GEMBA score has the same problem and is excluded from all reported statistics, but this does not affect the High-tier average shown here, since en-es is tiered Medium, not High (§4.2).

explanation we independently rebuilt Ukrainian’s quality scores from the raw WMT DA data and rejoined them to every metric checkpoint by segment ID: the alignment is correct, and the near-zero result is not a pipeline artifact. Investigation confirmed the cause: all 16,470 Ukrainian segments in the WMT DA dataset carry the domain label other (segment IDs: uk-en_2022_other), indicating non-news text, whereas every other language in our pipeline uses news-domain test sets. The short average sentence length (≈ 12 words) is consistent with a conversational or social-media register (Kocmi et al., 2024). A single-reference metric evaluated against conversational MT cannot discriminate quality adequately. Ukrainian should be treated as a domain outlier, not a language/script effect, consistent with Zouhar et al. (2024), who show that fine-tuned metrics degrade substantially under domain shift. We applied the same segment-ID rebuild-and-rejoin check to is-en and ha-en, the two other languages with weak aggregate correla-

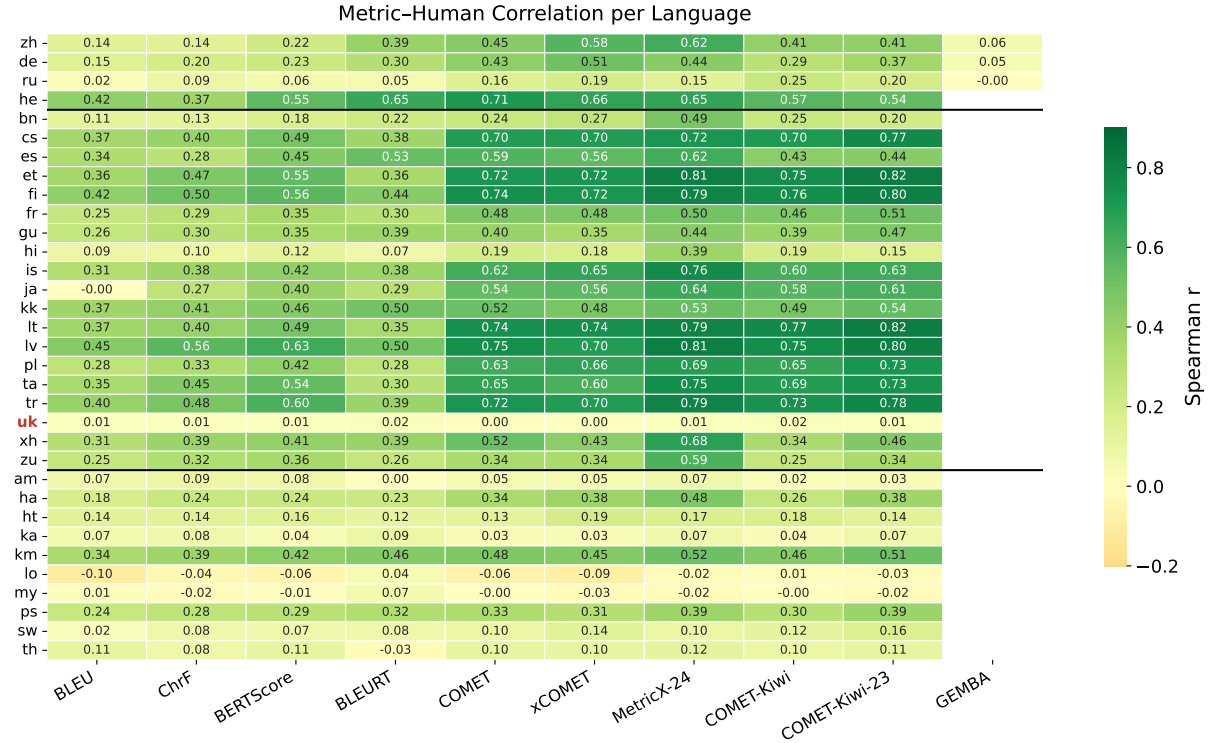


Figure 2: Per-language Spearman ρ heatmap for all evaluated metrics. Languages sorted by resource tier. Black horizontal lines divide tiers. Ukrainian (**uk**) is a domain outlier (non-news register; see §4.2). *Tier 2 languages* (am, ht, ka, lo, my, sw, th) use synthetic MQM annotations generated by Llama-3.1-8B; their correlations are exploratory and unvalidated against human judgments. GEMBA’s column is sparse by design: it is reported only for the high-resource MQM languages (§3.1), excluded from Tier 1b because its backbone was not fine-tuned on DA-style judgments, and excluded from Tier 2 to avoid circular evaluation with the Llama-3.1-8B-generated hypotheses. he-en is additionally blank for GEMBA specifically: its scores predate a segment-identifier correction applied to every other metric (§4.2) and could not be realigned (see Limitations).

tions (Table 8); both alignments are correct, and their weakness is annotation noise rather than a domain or pipeline effect.

Medium > High: Averaged across metrics, medium-resource languages outperform high-resource (COMET: $\rho = .526$ vs. $.436$; MetricX: $.588$ vs. $.465$; Figure 4 plots this for all ten metrics). This reverses the intuitive ordering and mirrors the Hindi anomaly reported by Yari et al. (2025) in Indian-language MT evaluation. The mechanistic explanation is that high-resource MQM languages span multiple WMT years (2020–23) and MT systems, introducing inter-year domain and system variance that suppresses aggregate correlation, a source of noise absent from the medium-tier DA languages.

Spanish (en-es WMT24) is classified in the medium tier despite using MQM annotation because it comes from a single year with a more recent system generation and annotation cohort. Had it been grouped with High, the medium>high gap

would be smaller, not larger: High would rise from $.436$ to approximately $.466$, and removing Spanish from Medium would lower the medium-tier mean; the ordering survives but with a narrower margin.

Russian anomaly, Simpson’s Paradox: Russian achieves the lowest MQM correlation of any non-domain-outlier language (aggregate COMET $\rho = .157$). Year-stratified analysis reveals the mechanism: WMT21-only yields $\rho = .383$ ($n = 15,600$) and WMT22-only yields $\rho = .501$ ($n = 19,725$), but pooling both years collapses to $.157$. This is a textbook aggregation bias: the two WMT system pools have different quality distributions, so across-year pooling creates a between-year confound that suppresses the within-year metric–human relationship. German shows the same pattern but less severely (2020: $\rho = .497$; 2021: $.323$; 2022: $.456$; aggregate: $.425$). The Russian aggregate is extreme because its two yearly pools are more divergent. This is not a property of Russian as a language, it is an artefact of pooling heteroge-

neous MT system generations.

Khmer: Despite being classified as low-resource, Khmer achieves COMET $\rho = .476$ and MetricX $\rho = .516$, higher than several medium-tier languages (e.g., Japanese .279, Hindi .194). Khmer is Austroasiatic (Mon-Khmer) and typologically *isolating* (not agglutinative as sometimes assumed), sharing that property with Chinese and Thai. Its strong performance here reflects single-year Wikipedia-domain data, which produces cleaner and more consistently-structured reference translations than news. This suggests that annotation consistency and domain homogeneity can compensate for low absolute resource level.

Hebrew shows no metric-family anomaly: COMET (reference-based) achieves the strongest correlation of any metric reported for this language ($\rho = .713$), ahead of xCOMET ($\rho = .660$), MetricX ($\rho = .653$), CK23 ($\rho = .540$), COMET-Kiwi-22 ($\rho = .569$), and BERTScore ($\rho = .552$), despite Hebrew having 12,441 professionally annotated segments. Figure 3 shows this COMET-over-CK23 gap is the largest deviation from the reference-free/reference-based equal-performance diagonal across all evaluated languages. An earlier version of this analysis reported the reverse pattern (reference-based COMET substantially underperforming reference-free QE) and attributed it to Hebrew being out-of-distribution for COMET-22’s training window. That result traced to a reference-omission error in the data-loading pipeline for the he-en span-fallback pair (see Limitations), not to a property of the metric or the language; correcting it reverses the finding. We validated the correction externally, not just internally, and at segment level rather than system level: aggregating to one mean per system before correlating (as an earlier draft did) is a weak check, since averaging cancels exactly the kind of per-segment misalignment this correction fixes. For both affected language pairs (he-en, en-es) and all four affected metrics (COMET, BLEURT, xCOMET, MetricX), we paired our corrected per-segment scores directly against the official mt-metrics-eval-v2 (Freitag et al., 2023) per-(system, segment) baseline releases for WMT23 he-en and WMT24 en-es, across all 8 (language, metric) checks ($n \approx 8,700$ – $8,900$ segment pairs each). COMET and xCOMET, whose checkpoints match the official baseline exactly (wmt22-comet-da, xCOMET-XL), agree closely with the official scores (Pearson/Spearman $r \geq 0.867/0.837$). BLEURT

and MetricX agree more weakly overall (Pearson/Spearman $r \geq 0.572/0.623$), which traces to a genuine checkpoint-version mismatch rather than a residual alignment error: we score BLEURT with an unofficial HF port (Elron/bleurt-large-512) rather than Google’s official BLEURT-20 checkpoint, and for he-en specifically the official WMT23 baseline is MetricX-23 while our pipeline scores every language with the newer MetricX-24.

Spanish (WMT24 en-es) achieves the strongest correlations of any MQM language: MetricX $\rho = .625$, COMET $\rho = .586$. Classified in the medium tier (single-year, recent system generation), it is consistent with en→es being the most resource-rich, morphologically regular fusional en→X pair in the benchmark: alphabetic script, abundant training data, and single-year WMT24 annotations with high inter-annotator agreement. Spanish confirms the upper bound of what learned metrics can achieve on well-matched conditions.

4.3 Statistical Robustness

Bootstrap 95% CIs are narrow for large-corpus languages (de COMET: [.418, .433], $n = 49,741$) and wider for small ones (ps COMET: [.307, .360], $n = 4,611$), confirming that high-confidence claims require $n \gtrsim 10,000$ segments. All Tier 1a and Tier 1b COMET correlations except Ukrainian are significant at $p < 10^{-10}$ or better. Ukrainian COMET $p = 0.90$ (domain outlier). Russian despite $n = 35,325$ achieves $p \ll 10^{-10}$, but $\rho = .157$, statistically significant but practically weak, highlighting that statistical significance and practical reliability are distinct in large- n meta-evaluation.

4.4 Translation Direction $\times$ Morphological Type

Table 2 reports the 2×3 cell means for COMET (full results across all ten metrics are in the released code and data repository). 12 languages contribute data in both translation directions; we separate them by segment-id prefix before computing cell means (per-language breakdowns in Table 8). The dominant pattern is directional: en→X languages score substantially above X→en across both morphological types.

- **Agglutinative:** en→X $\rho = 0.675$ vs. X→en $\rho = 0.440$, a gap of +0.235.
- **Fusional:** en→X $\rho = 0.563$ vs. X→en $\rho = 0.292$ (depressed by the Ukrainian domain

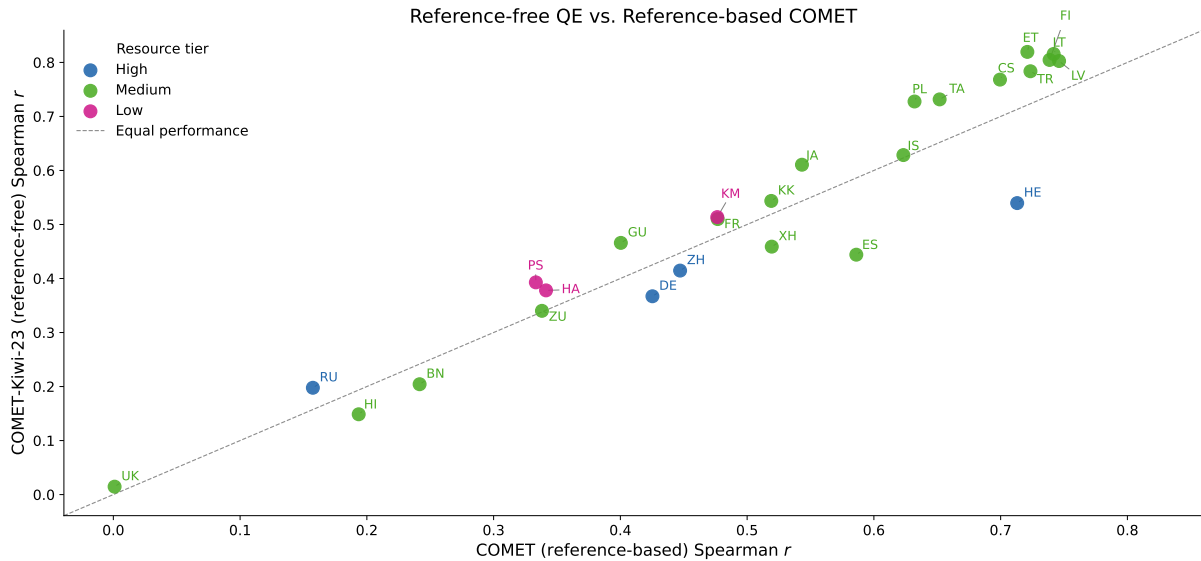


Figure 3: CK23 (reference-free) vs. COMET (reference-based) Spearman ρ for all 26 languages with human annotations. Most languages cluster near the equal-performance diagonal, with reference-free QE holding a modest overall edge. Hebrew and Spanish are the two largest deviations from the diagonal in the dataset, both in COMET’s favor (-0.173 and -0.142 Spearman points, CK23 minus COMET); we have no validated explanation for either and do not speculate one. uk sits below the diagonal for both metrics due to domain mismatch (see §4.2); ru reflects a Simpson’s Paradox aggregate.

outlier, included here for transparency rather than excluded), a gap of $+0.271$.

The difference-in-differences between the fusional direction gap and the agglutinative direction gap is 0.036 (Table 2), indicating the direction effect is largely constant across morphological types with only a modest interaction. The best metric varies by direction: CK23 leads $X \rightarrow \text{en}$ while MetricX leads $\text{en} \rightarrow X$ in 4 of 6 agglutinative languages.

Robustness: mixed-effects model. The one-way ANOVAs above treat every (language, direction) row as independent, but 12 languages contribute two correlated rows each (one per direction). To check whether this affects the conclusions, we fit a linear mixed-effects model via restricted maximum likelihood (REML): $\rho \sim \text{direction} + \text{morphology}$, random intercept per language, $n = 38$ observations, 26 languages. Direction remains highly significant ($X \rightarrow \text{en}$ coefficient -0.286 , $p = 7.5 \times 10^{-17}$); the fusional-vs-agglutinative morphology coefficient does not reach significance (-0.136 , $p = 0.118$). The language random intercept absorbs 0.79 of total variance (the intraclass correlation, ICC), confirming the non-independence this model was designed to check for. The mixed-effects model agrees with the ANOVA: morphological type is not an inde-

pendently detectable main effect once repeated-measures structure is modeled explicitly, and direction remains the dominant, robustly significant predictor.

Robustness: year stratification. Only de, ru, zh, and cs are multi-year in this cell analysis; every other contributing language is single-year. de and ru enter the $\text{en} \rightarrow X \times \text{fusional}$ cell via their pooled (not year-stratified) rho, and §5.2 already shows ru’s pooled value is Simpson’s-Paradox-suppressed. Substituting each language’s mean-of-per-year rho for its pooled value moves that cell’s mean from 0.563 to 0.591 , a small increase, meaning year-stratification if anything slightly strengthens the $\text{en} \rightarrow X$ advantage rather than explaining it away. A fully rigorous per-metric training-coverage control (checking whether each metric’s training-data volume differs by direction) is not attempted here: none of the 10 metrics publish per-language-pair-and-direction training coverage at the granularity this would require, and we flag it as future work.

Contrast with Mondshine et al. (2025): in their summarization setting, COMET correlates *better* for fusional languages (coherence Pearson ≈ 0.35 for high-fusional vs. 0.07 for isolating). In the MT setting the dominant factor is direction, not morphological type: fusional $X \rightarrow \text{en}$ languages score only $\rho = 0.292$ (depressed by the Ukrainian outlier),

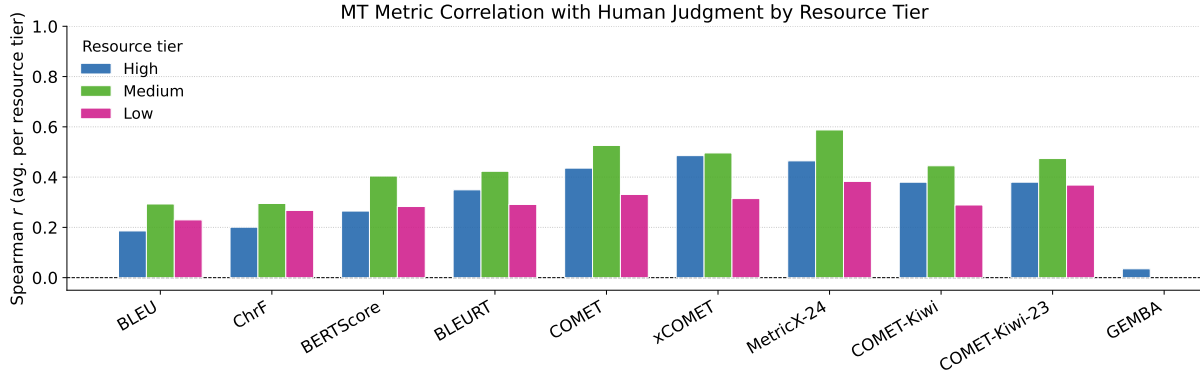


Figure 4: Average Spearman ρ by metric and resource tier. The medium $>$ high anomaly holds for all ten metrics with a Medium-tier value (GEMBA has none, excluded from Tier 1b by design: Table 1); it is explained by cross-year pooling variance in the high-resource MQM tier (see §4.2).

Dir.	Isol.	Agglutin.	Fusional
		Mean COMET ρ [95% CI]	
X $\rightarrow$ en	0.462 [—]	0.440 [0.313,0.538]	0.292[0.156,0.438]
en $\rightarrow$ X	—	0.675[0.619,0.720]	0.563[0.446,0.657]
X $\leftrightarrow$ Y	—	—	0.354[0.242,0.466]
η_{dir}^2		0.309	
η_{morph}^2		0.073	
η_{cell}^2		0.618	

Table 2: Mean COMET Spearman ρ by translation direction and morphological type. Cells with $n \leq 2$ languages have no CI (marked [—]). Empty (—) = no Tier-1 languages in that cell. η^2 via one-way ANOVA on per-language ρ . Russian ($\rho = .157$, en $\rightarrow$ X+fusional) and Ukrainian ($\rho = .001$, X $\rightarrow$ en+fusional) are included in their cells; see §4.2. **Bold** marks the X $\rightarrow$ en agglutinative cell specifically because it is where the reference-free metric CK23 (not shown in this COMET-only table) wins all 6/6 direction-separated pairs (Table 8); no other cell in this table carries that meaning, and η^2 values below are never bolded.

while fusional en $\rightarrow$ X languages score $\rho = 0.563$, a gap of +0.271 attributable to translation direction. Agglutinative languages show the same directional pattern ($\rho = 0.440$ X $\rightarrow$ en vs. $\rho = 0.675$ en $\rightarrow$ X, a gap of +0.235), confirming that direction dominates in the MT setting: within this design we cannot detect an independent morphological main effect (the empty en $\rightarrow$ X isolating cell limits what the data can rule out; see Limitations). $\eta_{\text{cell}}^2 = 0.618$ quantifies the combined effect; $\eta_{\text{dir}}^2 = 0.309$ isolates direction as the dominant contributor.

4.5 Typological Predictors of Metric Choice

Translation direction is the dominant predictor of metric reliability ($\eta_{\text{dir}}^2 = 0.309$; §3.3); we now ask whether that directional signal extends to metric

Direction	BLEU	ChrF++	COMET	xCOMET	MetricX
en $\rightarrow$ X	0.303	0.380	0.605	0.602	0.671
X $\rightarrow$ en	0.230	0.253	0.364	0.350	0.392
X $\leftrightarrow$ Y	0.202	0.247	0.354	0.340	0.530

Table 3: Mean Spearman ρ by translation direction for five representative metrics (averaged across all languages in each direction group). MetricX-24 leads across all three directions among these five metrics, including X $\rightarrow$ en, where it now edges out CK23 (not shown) by a narrow margin ($\rho=0.392$ vs. $\rho=0.387$); this pooled ranking is not the same question as which metric wins the most individual X $\rightarrow$ en language pairs (Table 4), where CK23 remains the more frequent winner. X $\leftrightarrow$ Y pairs show the lowest surface-metric scores but near-parity on MetricX.

choice, that is, which metric a practitioner should prefer for an unseen language.

For each of the 26 languages with human annotations, we rank all metrics by Spearman ρ and record the best-performing metric. Table 4 summarizes these 35 direction-language pair observations (34 in the main pipeline; de-en held out as a blind validation), with leave-one-out (LOO) cross-validation for cells with $n \geq 6$ (overall 19/26 = 73%). Isolating X $\rightarrow$ en ($n=2$) and X $\leftrightarrow$ Y ($n=5$) fall below this threshold and are marked ‘—’ in the LOO column rather than validated. The most robust pattern in the table needs no morphological breakdown at all: CK23 beats COMET in all 6 of 6 agglutinative and 7 of 7 fusional en $\rightarrow$ X pairs (13/13 overall), both cells with bootstrap CI [1.00, 1.00] on the win fraction. Table 4 is organized first by direction, the dominant factor, and then by morphological type within each direction as a secondary refinement: even though morphology explains substantially less

of the aggregate variance in reliability than direction does ($\eta_{\text{morph}}^2 = 0.073$ vs. $\eta_{\text{dir}}^2 = 0.309$), the within-direction agglutinative-vs-fusional splits are large enough to flip which specific metric wins the discrete argmax comparison (§3.3), so morphology remains a useful secondary key for the practitioner-facing recommendation even though it is not the primary driver of reliability. Three patterns emerge:

1. **X→en: CK23 wins consistently across morphological types.** CK23 is the best metric for agglutinative X→en (6/6, LOO 6/6; bootstrap CI on CK23>COMET: [1.00, 1.00]) and fusional X→en (6/9 including Hebrew and de-en held-out, LOO 5/7 over the smaller DA-pipeline-only subset; see Table 4’s note). De-en (German→English, $n=134,792$) is genuinely unseen, German enters the pipeline via WMT MQM en→de only; CK23 achieves $\rho=.414$ vs. COMET $\rho=.383$ and MetricX $\rho=.389$, confirming the prediction. The reference-free design sidesteps the n -gram overlap problem that BLEU and ChrF++ face when rich morphology produces sparse token matches.
2. **en→X: MetricX competitive for agglutinative; CK23 leads for fusional.** For agglutinative en→X (6 languages), MetricX wins 4/6 (LOO 4/6), though margins are often small (kk: MetricX 0.770 vs. CK23 0.766; tr: 0.787 vs. 0.784). CK23 still beats COMET in 6/6 agglutinative en→X pairs (bootstrap CI [1.00, 1.00]), so the CK23-vs-COMET choice is unambiguous; the MetricX advantage only matters when choosing between all three neural metrics. For fusional en→X (7 languages), CK23 wins 4/7 and beats COMET in 7/7 (CI [1.00, 1.00]).
3. **Isolating X→en and X↔Y: MetricX.** MetricX-24 wins both isolating X→en cases (zh: +0.478 over BLEU; km: +0.174; average +0.326) and dominates X↔Y pairs (4/5), where neither source nor target is English and reference-free QE encoders are off-domain.

Across all profiles, the gain over BLEU from choosing the recommended metric ranges from +0.130 to +0.432 Spearman ρ (Table 4), the “always use a neural metric” message is robust even where the *choice* among neural metrics is uncertain.

As a post-hoc check, we trained a Ridge regression predictor (World Atlas of Language Structures (WALS; Dryer and Haspelmath, 2013) typological features + morphological type + direction,

leave-one-language-out cross-validation (LOLO-CV), $n=26$ Tier-1 languages). Pearson $r = 0.47$ for CK23 and $r = 0.56$ for Kiwi-22; recommendation accuracy is 5/26, too low to treat the predictor as deployable. The moderate r for the CK family confirms that direction and CK advantage are partially co-predictable from typological features, but the low recommendation accuracy is a negative result on the stronger claim that typological features alone suffice to select among neural metrics.

5 Discussion

5.1 The Neural-Surface Gap Scales With Morphological Complexity

This is a claim about a different quantity than §3.3: $\eta_{\text{morph}}^2 = 0.073$ measures how much morphological type explains variance in absolute metric–human correlation, and it is small. The finding here concerns the *gap* between neural and surface metrics, a derived quantity that morphology predicts much more strongly than it predicts either metric’s raw reliability. COMET, xCOMET, and MetricX-24 dominate at every tier and in every typological cell (Tables 1–3), but that gap over surface metrics is not uniform: it is largest for morphologically complex languages and non-news domains, precisely the conditions where evaluation is most needed and most difficult.

For fusional X→en languages (excluding the Ukrainian outlier), the COMET–BLEU gap averages 0.119 Spearman points; holding direction fixed, agglutinative X→en languages show a larger gap still, 0.151 (the same pattern holds en→X, where the gap is 0.351 for agglutinative languages). This is not simply a ceiling artifact of agglutinative X→en being an easier cell overall: within X→en, BLEU’s own reliability is nearly flat across morphological type ($\rho \approx 0.230$ pooled; agglutinative and fusional differ by only a few points), while COMET’s reliability rises substantially from fusional to agglutinative. The widening gap comes from COMET tracking something morphology-related that BLEU does not, not from BLEU degrading uniformly.

This is consistent with n -gram overlap systematically underestimating adequacy when target morphology is rich: a correct translation of a Finnish verb may share zero surface n -grams with the reference if a different but equally valid inflected form was chosen. Neural metrics are comparatively insensitive to this effect, which is exactly why mor-

Direction	Morphological type	n pairs	Recommended metric	LOO accuracy	Avg $\Delta\rho$ vs. BLEU
X→en	Agglutinative	6	CometKiwi-23 (6/6)	6/6	+0.196
X→en	Isolating	2	MetricX-24 (2/2)	–	+0.326
X→en	Fusional	9 [†]	CometKiwi-23 (6/9)	5/7	+0.130
en→X	Agglutinative	6	MetricX-24 (4/6) [‡]	4/6	+0.432
en→X	Fusional	7	CometKiwi-23 (4/7)	4/7	+0.356
X↔Y	Fusional	5	MetricX-24 (4/5)	–	+0.327

Table 4: **Typological metric selection guide:** empirically observed best metric per profile across 35 direction-language pair observations (de-en held out). n = number of direction-separated pairs; LOO accuracy = leave-one-out cross-validation (– = cell too small to validate); Avg $\Delta\rho$ = mean Spearman gain over BLEU. [†]Fusional X→en’s $n=9$ includes Hebrew and the held-out de-en pair, neither scored by the DA direction-separated pipeline the LOO cross-validation runs over; its LOO column (5/7) therefore validates only the 7 DA-pipeline languages, a smaller denominator than the win count. Held-out de-en ($n=134,792$, absent from main pipeline): CometKiwi-23 $\rho=.414$ vs. COMET $\rho=.383$, confirming prediction. [‡]Agglutinative en→X: MetricX-24 wins 4/6, but CometKiwi-23 beats COMET in 6/6 (bootstrap CI [1.00, 1.00]); advantage often $< 0.004 \rho$ (kk, tr).

phology moves their reliability much less than it moves direction ($\eta_{\text{morph}}^2 = 0.073$ vs. $\eta_{\text{dir}}^2 = 0.309$) while still moving the size of their advantage over BLEU. We note this pattern rests on two X→en cells without bootstrap CIs on the gap itself, and should be read as a documented pattern in the available data rather than a strongly powered general law.

5.2 Domain Is the Confounding Variable, Not Resource Level

The Ukrainian domain confound provides a clean natural experiment: among X→en direction-controlled languages using news or Wikipedia domains (excluding ha-en and is-en, whose near-zero correlations across all metrics are attributed to annotation noise rather than domain effects (see Table 8)), COMET ρ ranges from approximately .144 (ja-en) to .613; Ukrainian (domain other) sits at $\rho = .001$, 0.143 below the next-lowest X→en language (ja-en, $\rho \approx .144$). Kocmi et al. (2024) note WMT24 comprised “news data and four other separate domains”, precisely the inconsistency creating the Ukrainian outlier. Zouhar et al. (2024) confirm that fine-tuned metrics degrade substantially under domain shift. The practical implication is that the diagnosis “our language is low-resource, so metrics are unreliable” is often wrong in a correctable way: resource level and domain mismatch require different remedies.

This fits the broader pattern established in §3.3: translation direction is the dominant predictor of metric reliability ($\eta_{\text{dir}}^2 = 0.309$; Table 2), accounting for nearly a third of between-language variance, while morphological type is a substantially smaller main effect ($\eta_{\text{morph}}^2 = 0.073$); the joint

direction×morphology cells ($\eta_{\text{cell}}^2 = 0.618$) capture additional within-direction morphological variation, particularly the agglutinative vs. fusional split within X→en ($\rho = 0.440$ vs. $\rho = 0.292$). Domain mismatch and translation direction, not resource level, are the two factors that actually predict where metrics fail.

Always stratify by year. Table 5 documents two failure modes from multi-year pooling: *suppression* (Russian pooled $\rho = .157$ hides within-year [0.38, 0.50], $\Delta = +0.285$, from heterogeneous WMT21/22 system generations) and *masking of heterogeneity* (Czech cs-en pooled $\rho = .288$ masks a per-year range of [−0.01, 0.69], including complete metric failure in 2021 and strong correlation in 2017). Both are correctable by year-stratification.

Lang	Tier	Years	Pooled ρ	Year ρ range	Δ
ru	MQM	2021–22	.157	[0.38, 0.50]	+0.285
zh	MQM	2020–22	.447	[0.47, 0.57]	+0.054
de	MQM	2020–22	.425	[0.32, 0.50]	+0.000
cs-en	DA	2017–22	.288	[−0.01, 0.69]	+0.012

Table 5: Year-stratified vs. pooled COMET Spearman ρ for languages with multiple WMT years. Δ = stratified mean – pooled. Year ρ range = [min, max] per-year ρ . Russian (suppression: $\Delta = +0.285$) and Czech cs-en (masking: pooled $\approx .288$, but year range [−0.01, 0.69]) show two distinct aggregation failure modes.

6 Conclusion

We presented a unified meta-evaluation of ten MT metrics across 33 languages, three annotation tiers, and 730,748 segments, with the goal of understanding *when* metrics can be trusted and *why* they fail.

The central finding is that metric failure correlates with domain mismatch and the joint configuration of translation direction and morphological type far more strongly than with resource level. Ukrainian’s catastrophic failure is entirely explained by a genre confound; Russian’s weak aggregate correlation is a Simpson’s Paradox artefact from multi-year pooling. Neither finding is about language difficulty, both are correctable evaluation design choices.

Translation direction is the largest quantified predictor of metric reliability in our study ($\eta_{\text{dir}}^2 = 0.309$; Table 2): $\text{en} \rightarrow \text{X}$ is reliably better-evaluated than $\text{X} \rightarrow \text{en}$ regardless of morphological type. Morphological type is a substantially smaller main effect ($\eta_{\text{morph}}^2 = 0.073$), though within $\text{X} \rightarrow \text{en}$, agglutinative languages ($\rho \approx 0.440$) outperform fusional ones ($\rho \approx 0.292$). The practical implication is concrete: deploying COMET or MetricX to evaluate $\text{en} \rightarrow \text{X}$ translation into agglutinative targets should be expected to be reliable ($\rho \approx 0.675$); deploying the same metrics for $\text{X} \rightarrow \text{en}$ fusional targets with mixed-year data should not ($\rho \approx 0.292$, and even lower when pooling artefacts are present).

Neural metrics (MetricX-24, xCOMET, COMET) are substantially better than surface metrics in exactly the settings where surface metrics are most relied upon: the $\text{en} \rightarrow \text{X}$ gap between MetricX-24 ($\rho \approx 0.671$) and BLEU ($\rho \approx 0.303$) is large enough that any low-resource evaluation still using BLEU as its primary metric should be revisited.

A practical decision procedure. Table 4 reduces to a short recipe for choosing a metric on an unseen language pair:

1. **Always prefer a neural metric over BLEU/ChrF++.** The gap ranges from +0.130 to +0.432 Spearman points and only widens for morphologically complex or non-news-domain languages (§5).
2. **$\text{X} \rightarrow \text{en}$, agglutinative or fusional target: use CometKiwi-23.** It wins these two $\text{X} \rightarrow \text{en}$ cells and beats COMET specifically in 6/6 agglutinative and 6/9 fusional pairs. (Isolating $\text{X} \rightarrow \text{en}$ is a separate case: see below.)
3. **$\text{en} \rightarrow \text{X}$, fusional target: use CometKiwi-23. $\text{en} \rightarrow \text{X}$, agglutinative target: MetricX-24 usually edges it out,** but the choice barely matters for the reference-free-vs-reference-based question: CometKiwi-23 beats COMET

in 13/13 $\text{en} \rightarrow \text{X}$ pairs regardless of morphological type.

4. **Isolating $\text{X} \rightarrow \text{en}$ or $\text{X} \leftrightarrow \text{Y}$ pairs: use MetricX-24.**

5. **Typological features alone will not pick the metric for you:** our Ridge predictor over WALS (Dryer and Haspelmath, 2013) features reaches only 5/26 recommendation accuracy. Direction and morphology as defined here are the operative signal, not abstract typology.

Future work should validate Tier 2 synthetic annotations against human judgments, extend the $\text{direction} \times \text{morphology}$ analysis to polysynthetic languages, and determine *why* direction predicts reliability rather than only that it does. Our claim throughout is predictive/associational (direction is the dominant term in an ANOVA and a mixed-effects model), not causal. Distinguishing candidate mechanisms (asymmetric training-data coverage by direction, source-side complexity, or reference-generation artefacts specific to translating into vs. out of English) would require controlled intervention on training data or synthetic directional pairs, not further observational analysis of existing WMT releases.

Broader coverage is the most direct extension: the analysis pipeline is language-agnostic and will incorporate new WMT DA pairs as they are released. The key open question is whether the $\text{direction} \times \text{morphology}$ cell structure holds for polysynthetic languages (e.g. Inuktitut) and for the currently empty abjad $\times$ agglutinative cell, which Arabic would fill.

A within-language, within-year MQM-versus-DA comparison on en-de, zh-en, and en-ru would isolate the annotation-protocol effect from resource level and training-data coverage; if the medium-tier advantage collapses under matched protocol, §4 requires reinterpretation.

Limitations

Tier 2 synthetic annotations have not been validated against human judgments; Tier 2 results are exploratory. The Ukrainian result is a domain confound (uk-en_2022_other), not a linguistic finding. Russian and German aggregate correlations are depressed by multi-year system-pool heterogeneity (Simpson’s Paradox) and should not be interpreted as language-level metric failure. We evaluate he-en and en-es against a single reference

(refA) throughout; WMT23 he-en additionally provides a second official reference (refB), not used in this analysis, so a multi-reference sensitivity check remains future work. GEMBA’s he-en and en-es scores are excluded from all reported statistics: they were computed before the segment-identifier fix described above and cannot be safely realigned to the corrected data (a positional remapping was attempted and rejected after it produced an implausible near-zero correlation against COMET, versus $\rho \geq 0.77$ segment-level agreement between two independently-correct metrics, COMET and xCOMET, on the same languages). Of these, only he-en is a Tier 1a (High-resource) language, so Table 1’s High-tier GEMBA value averages only de/ru/zh, not all four Tier 1a languages; en-es is tiered Medium and GEMBA is not reported for Medium/Low at all (Table 1’s caption), so its exclusion has no separate effect on any reported average.

Four directional variants present in the WMT DA dataset were not included in the main pipeline (de-en, ru-en, de-cs, fr-de); incorporating these would add fusional $X \rightarrow en$ and $X \leftrightarrow Y$ observations to the $direction \times morphology$ analysis.

Empty $en \rightarrow X$ isolating cell. All isolating languages in the benchmark (zh, km, th, lo, my) are $X \rightarrow en$ or Tier 2; there are no Tier-1 $en \rightarrow X$ isolating observations. The relatively small $\eta_{morph}^2 = 0.073$ may partly reflect this structural gap: if $en \rightarrow X$ isolating languages had been available, the morphological main effect might have been larger.

Language non-independence in the $direction \times morphology$ analysis. 12 languages contribute per-language ρ values in both translation directions; these observations are not independent, and the correct specification is a mixed-effects ANOVA with language as a random factor. We report this model in §3.3: $direction$ remains significant, $morphology$ does not. The one-way ANOVA is retained alongside it because it is what the $direction \times morphology$ cell means and their bootstrap confidence intervals are computed from.

Annotation protocol and resource tier are perfectly confounded. Every Tier 1a (high-resource) language uses MQM annotation; every Tier 1b (medium/low-resource) language uses DA. This means the “medium > high” pattern and all tier-level metric differences have an alternative explanation: annotation protocol, not resource level. MQM annotators are trained to flag subtle adequacy and fluency errors that DA crowd-workers routinely

smooth over, which compresses metric–human correlation independently of anything typological. We cannot separate these effects with the current data, and a dedicated study comparing MQM-annotated and DA-annotated pairs for the *same language* would be required to attribute the effect causally. Readers should treat tier-level comparisons as confounded on this axis.

A partial, protocol-controlled check is possible within Tier 1b alone, where medium and low resource levels are both DA-annotated: the raw gap (medium 0.441 vs. low 0.312) narrows once Hausa, the same $X \rightarrow en$ outlier already flagged as likely annotation noise in Table 8, is excluded (medium 0.441 vs. low 0.388), leaving only two low-resource languages and no reliable residual gap. This does not resolve the Tier 1a-vs-Tier 1b confound, which remains unfixable without a language annotated under both protocols, but it is evidence against resource level driving the pattern in the one place protocol can be held constant.

Ethics Statement

This work is a meta-evaluation of existing MT metrics using publicly available datasets (WMT MQM, WMT DA, FLORES+) and does not collect new human annotations or deploy systems in production. The Tier 2 synthetic annotations are generated by prompting Llama-3.1-8B-Instruct, a publicly released model, and are used only for exploratory correlation analysis clearly labelled as unvalidated. We explicitly flag synthetic Tier 2 results as exploratory throughout and exclude GEMBA from that tier to avoid circular evaluation.

No personally identifiable information is used or generated. The benchmark languages include several low-resource and endangered languages (Hausa, Khmer, Pashto, Swahili, and others); we make no claims about translation quality in those languages beyond the statistical patterns observable in the available test sets, and we caution that metrics should not be used for high-stakes decisions in settings where they have not been separately validated on domain- and script-appropriate data.

Acknowledgments

Compute resources were provided by the BYU Office of Research Computing. This work was supported by the BYU Computer Science Graduate Fellowship and generous donations to the BYU Machine Translation Research and Interlingual

eXperimentation (MATRIX) Lab, run by Dr. Steve Richardson and Dr. Eric Ringger. Thanks to my advisor, Dr. Eric Ringger, for his guidance and for proofreading this paper. This project grew out of a course taught by Dr. Nancy Fulda and Jay Orten. We also thank the anonymous WMT reviewers for their valuable feedback.

References

- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. [Findings of the 2017 conference on machine translation \(WMT17\)](#). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.
- Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. [It’s easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649.
- Daniel Deutsch, George Foster, and Markus Freitag. 2023. [Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. [WALS Online](#). Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). In *Transactions of the Association for Computational Linguistics*, volume 9, pages 1460–1474.
- Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. [Results of the WMT23 metrics shared task: Metrics might be guilty but references are not innocent](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628.
- 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](#). In *Transactions of the Association for Computational Linguistics*, volume 10, pages 522–538.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. [Can machine translation systems be evaluated by the crowd alone](#). *Natural Language Engineering*, 23(1):3–30.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). In *Transactions of the Association for Computational Linguistics*, volume 12, pages 979–995.
- 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.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024. [Findings of the WMT 2024 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.
- Tom Kocmi and Christian Federmann. 2023a. [GEMBA-MQM: Detecting translation quality error spans with GPT-4](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775.
- Tom Kocmi and Christian Federmann. 2023b. [Large language models are state-of-the-art evaluators of translation quality](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203.
- Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. [Beyond N-grams: Rethinking evaluation metrics and strategies for multilingual abstractive summarization](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19019–19035.
- NLLB Team and Marta R. Costa-jussà. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- 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.

- 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. 2022a. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.
- Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. [Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.
- Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. [Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234.
- Evan James Williams. 1959. [The comparison of regression variables](#). *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):396–399.
- Amir Hossein Yari, Kalmit Kulkarni, Ahmad Raza Khan, and Fajri Koto. 2025. [Revisiting metric reliability for fine-grained evaluation of MT and summarization in Indian languages](#). *arXiv preprint arXiv:2510.07061*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations*.
- 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.

Code	Language	Family	Script	Morphology	Resource	Annotation
am	Amharic [†]	Afro-Asiatic	Abugida	Agglutinative	Low	Synthetic MQM
bn	Bengali	Indo-European	Abugida	Fusional	Medium	DA
cs	Czech	Indo-European	Alphabetic	Fusional	Medium	DA
de	German	Indo-European	Alphabetic	Fusional	High	MQM
es	Spanish	Indo-European	Alphabetic	Fusional	Medium	MQM
et	Estonian	Uralic	Alphabetic	Agglutinative	Medium	DA
fi	Finnish	Uralic	Alphabetic	Agglutinative	Medium	DA
fr	French	Indo-European	Alphabetic	Fusional	Medium	DA
gu	Gujarati	Indo-European	Abugida	Fusional	Medium	DA
ha	Hausa	Afro-Asiatic	Alphabetic	Fusional	Low	DA
he	Hebrew	Afro-Asiatic	Abjad	Fusional	High	MQM
hi	Hindi	Indo-European	Abugida	Fusional	Medium	DA
ht	Haitian Creole [†]	Creole	Alphabetic	Fusional	Low	Synthetic MQM
is	Icelandic	Indo-European	Alphabetic	Fusional	Medium	DA
ja	Japanese	Japonic	Logographic	Agglutinative	Medium	DA
ka	Georgian [†]	Kartvelian	Alphabetic	Agglutinative	Low	Synthetic MQM
kk	Kazakh	Turkic	Alphabetic	Agglutinative	Medium	DA
km	Khmer	Austroasiatic	Abugida	Isolating	Low	DA
lo	Lao [†]	Tai-Kadai	Abugida	Isolating	Low	Synthetic MQM
lt	Lithuanian	Indo-European	Alphabetic	Fusional	Medium	DA
lv	Latvian	Indo-European	Alphabetic	Fusional	Medium	DA
my	Burmese [†]	Sino-Tibetan	Abugida	Isolating	Low	Synthetic MQM
pl	Polish	Indo-European	Alphabetic	Fusional	Medium	DA
ps	Pashto	Indo-European	Abjad	Fusional	Low	DA
ru	Russian	Indo-European	Alphabetic	Fusional	High	MQM
sw	Swahili [†]	Niger-Congo	Alphabetic	Agglutinative	Low	Synthetic MQM
ta	Tamil	Dravidian	Abugida	Agglutinative	Medium	DA
th	Thai [†]	Tai-Kadai	Abugida	Isolating	Low	Synthetic MQM
tr	Turkish	Turkic	Alphabetic	Agglutinative	Medium	DA
uk	Ukrainian	Indo-European	Alphabetic	Fusional	Medium	DA
xh	Xhosa	Niger-Congo	Alphabetic	Fusional	Medium	DA
zh	Chinese	Sino-Tibetan	Logographic	Isolating	High	MQM
zu	Zulu	Niger-Congo	Alphabetic	Fusional	Medium	DA

Table 6: All 33 languages in the benchmark, with language family (§3.3), script type, morphological type, resource tier, and annotation protocol (MQM = professional Multidimensional Quality Metrics; DA = crowd-sourced Direct Assessment; Synthetic MQM = Tier 2, unvalidated). [†]Tier 2 languages have no human MT annotations (§3.1).

A Benchmark Languages

Table 6 lists all languages in the benchmark with their family, script, morphological type, resource tier, and annotation protocol, generated directly from the same classification dictionaries used throughout the analysis (§3.3) so it cannot drift from what the pipeline actually used.

B Per-Language Spearman ρ by Direction

Table 7 reports COMET and MetricX-24 Spearman ρ for every language in the benchmark, grouped by translation direction. Languages are sorted within each direction group by COMET ρ descending. This table supports the direction $\times$ morphology analysis in §3.3 and Table 2.

Lang	Morph	COMET ρ	MetricX ρ
<i>en→X (English source)</i>			
lv	fusional	0.746	0.807
lt	fusional	0.742	0.790
fi	agglutin.	0.739	0.788
tr	agglutin.	0.724	0.787
et	agglutin.	0.721	0.808
cs	fusional	0.699	0.723
ta	agglutin.	0.652	0.751
pl	fusional	0.632	0.691
is	fusional	0.623	0.759
es	fusional	0.586	0.625
ja	agglutin.	0.543	0.636
de	fusional	0.425	0.438
ha	fusional	0.341	0.479
ru	fusional	0.157	0.146
<i>X→en (English target)</i>			
he	fusional	0.713	0.653
kk	agglutin.	0.519	0.532
km	isolating	0.476	0.516
zh	isolating	0.447	0.621
gu	fusional	0.400	0.437
ps	fusional	0.333	0.388
uk	fusional	0.001	0.012
<i>X↔Y (neither language is English)</i>			
xh	fusional	0.519	0.680
fr	fusional	0.477	0.502
zu	fusional	0.338	0.593
bn	fusional	0.242	0.485
hi	fusional	0.194	0.388

Table 7: COMET and MetricX-24 Spearman ρ per language, grouped by translation direction and sorted by COMET ρ descending within each group. **Bold** = best between COMET and MetricX-24 per row; MetricX-24 wins all but two languages (ru, he), where COMET leads instead. lv/lt are fusional Baltic (not agglutinative); km is isolating Austroasiatic; he is Semitic (abjad script). Full 10-metric results are in the released code and data repository. The 12 languages with data in both directions (fi, tr, et, lt, cs, ja, kk, ta, ha, pl, is, gu) appear once, under whichever direction section this table places them in, using that direction’s own value, not a cross-direction pooled average; see Table 8 for both directions’ values side by side.

Lang	Dir	Morph	COMET	CK23	<i>n</i>
<i>Agglutinative X→en</i> (CK23 wins 6/6; best over BLEU)					
fi-en	X→en	agglutin.	0.525	0.578	46,145
et-en	X→en	agglutin.	0.565	0.619	20,496
tr-en	X→en	agglutin.	0.527	0.588	30,186
kk-en	X→en	agglutin.	0.519	0.544	6,789
ta-en	X→en	agglutin.	0.361	0.428	7,577
ja-en	X→en	agglutin.	0.144	0.156	46,747
<i>Agglutinative en→X</i> (MX wins 4/6; CK23>COMET in 6/6)					
en-fi	en→X	agglutin.	0.739	0.805	34,335
en-et	en→X	agglutin.	0.721	0.819	13,376
en-kk	en→X	agglutin.	0.674	0.766	8,219
en-ta	en→X	agglutin.	0.652	0.732	7,890
en-tr	en→X	agglutin.	0.724	0.784	5,171
en-ja	en→X	agglutin.	0.543	0.611	24,434
<i>Fusional X→en</i> (CK23 wins 6/9 including de-en held-out; LOO 5/7)					
lt-en	X→en	fusional	0.488	0.534	10,315
cs-en	X→en	fusional	0.288	0.344	51,097
pl-en	X→en	fusional	0.290	0.350	11,816
gu-en	X→en	fusional	0.400	0.466	9,063
ps-en	X→en	fusional	0.333	0.393	4,611
is-en	X→en	fusional	0.067	0.055	8,903
ha-en	X→en	fusional	0.045	0.036	13,171
de-en*	X→en	fusional	0.383	0.414	134,792
<i>Fusional en→X</i> (CK23 wins 4/7; CK23>COMET in 7/7)					
en-lt	en→X	fusional	0.742	0.816	8,959
en-cs	en→X	fusional	0.699	0.768	70,943
en-pl	en→X	fusional	0.632	0.728	10,572
en-gu	en→X	fusional	0.673	0.754	6,924
en-ha	en→X	fusional	0.341	0.378	10,812
en-is	en→X	fusional	0.623	0.628	10,838
en-lv	en→X	fusional	0.746	0.803	5,810

Table 8: Direction-separated COMET and CK23 Spearman ρ across 15 WMT DA languages. Direction parsed from segment-id prefix. Bold = best between COMET and CK23. $X \leftrightarrow Y$ -direction languages (bn, fr, hi, xh, zu; Table 4) are not direction-separated and are not shown here. For $en \rightarrow X$ agglutinative/fusional, MetricX-24 often wins: MX wins en-kk (0.770), en-ta (0.751), en-tr (0.787), en-ja (0.636), en-ha (0.479), en-is (0.759), en-lv (0.807). ha-en and is-en show near-zero correlations across all metrics, likely due to annotation noise rather than a typological pattern. *de-en (German→English) is a held-out pair absent from the main pipeline (German enters via WMT MQM $en \rightarrow de$); CK23 wins ($\rho=.414$ vs. COMET $\rho=.383$), confirming the fusional $X \rightarrow en$ prediction.