AQI: a composite human-calibrated metric for machine translation quality assessment

Anastasiya Serdioukova
Alconost
nastya@alconost.com

Denis Zhilko
Alconost
dz@alconost.com

Abstract

Manual evaluation of machine translation (MT) quality is the gold standard but expensive and noisy, and at production volumes it does not scale: automatic evaluation must carry the main load, with humans kept as a corrective signal on a limited subset. We audit six automatic metrics across three families—surface (chrF, TER), embedding (BERTScore F1), and learned (COMETKiwi-XXL, MetricX-XXL, xCOMET-XXL)—on the WMT25 General-MT corpus ($n = 3{,}450$ segment-system ratings, 5 language pairs (LPs); en→es drawn from WMT24, see §2.1). No single metric reaches the human-agreement ceiling: the strongest (MetricX-XXL) only approaches it, and combining metrics adds a small but reliable margin. Production stakeholders, meanwhile, expect a single scalar in $0-100$ with clear quality thresholds, not a vector of six heterogeneous numbers. We propose the **Aggregate Quality Index (AQI)**—a composite, human-calibrated index in two equal-billing versions differing in the normalisation of the key metric, with an optional per-LP calibration. AQI is interpretable and production-friendly: a single score, explicit weights, a transparent path from input metrics to the final number. We further derive a simple human-in-the-loop index—a linear blend of one human score into the auto-only AQI—and select the blend weight via honest validation against an independent annotator.

1 Introduction

1.1 The cost and noise of human evaluation

Human evaluation of MT quality produces **two kinds of signal**: error spans with severities—Multidimensional Quality Metrics (MQM; Lommel et al., 2014) and its lighter derivative, Error Span Annotation (ESA; Kocmi et al., 2024b)—and a **holistic score**, a Direct Assessment (DA) style 0–100 judgement of the whole segment. In WMT25 both come from the same ESA annotation: the annotator marks spans, then scores the segment. Both signals are expensive (trained annotators, linear scaling) and noisy: Freitag et al. (2021) and Freitag et al. (2023) document substantial inter-annotator disagreement in MQM, while Hendy et al. (2023) and Fischer and Läubli (2020) identify a systematic **AI bias**—annotators under-rate translations explicitly labelled as machine-produced. Averaging over the segment makes the holistic score **less noisy** than the additive span penalty, so we adopt it as the **primary oracle** for AQI calibration, with the span penalty as an input signal and a diagnostic channel (§2.2).

1.2 Automatic metrics: families and selection

Modern automatic translation evaluation is organised into three families: **surface-level** metrics based on lexical overlap, **embedding-based** metrics built on contextual representations, and **learned (neural)** metrics trained directly on human judgements. We compute **six representatives** across these families (Table 1)—chrF (Popović, 2015) and TER (Snover et al., 2006) (surface), BERTScore F1 (Zhang et al., 2020) (embedding), and the learned metrics COMETKiwi-XXL (Rei et al., 2022b, 2023), reference-based MetricX-XXL (Juraska et al., 2023) and xCOMET-XXL (Guerreiro et al., 2024). We deliberately **exclude BLEU** (Papineni et al., 2002): recent WMT meta-evaluations show its failure on neural MT (Mathur et al., 2020; Kocmi et al., 2021), and the WMT-22 Metrics Shared Task (Freitag et al., 2022) explicitly recommended "stop using BLEU", which our own production experience confirms. Per-metric definitions, scales, training signals and known failure modes are given in Appendix A; here we note only the orientations that matter downstream: MetricX-XXL is an error score on $[0, 25]$ (lower-is-better) and TER is lower-is-better, while the other four are higher-is-better.

Proceedings of the Eleventh Conference on Machine Translation, pages 654–672
October 28-29, 2026 ©2026 Association for Computational Linguistics

Family	Metric	Mode
Surface	chrF	ref
	TER	ref
Embedding	BERTScore F1	ref
Learned	MetricX-XXL	ref
	xCOMET-XXL	ref
	COMETKiwi-XXL	QE

Table 1: Taxonomy of automatic MT quality metrics. *ref*—reference-based; *QE*—quality estimation (reference-free).

The per-metric audit in §2.3 computes all six on the full WMT25 multi-domain corpus and reports their correlation with the human score and with each other, the area under the ROC curve (AUC) on major-error detection, and a redundancy analysis—which metrics add an independent signal and which can be dropped from a composite without loss.

1.3 Why AQI is needed

Three arguments motivate a composite index on top of the existing inventory: **scalability** (manual error annotation does not scale to production volumes, so automatic evaluation must bear the main load); **no single metric reaches the human ceiling** (§2.3: even the strongest metric only approaches inter-annotator agreement, and a combination buys a small but reliable margin); and **interpretability** (stakeholders require a single 0–100 scalar with clear thresholds, not a six-dimensional vector).

1.4 Related work

Combining automatic metrics is not new. RegEMT (Štefánik et al., 2021) fits a regression ensemble over surface, syntactic and semantic metrics and reports consistent gains over any standalone metric, but treats the resulting complexity as a cost to be managed rather than removed. COMET-22 (Rei et al., 2022a) likewise ensembles a DA-trained estimator with an MQM-trained multitask model, with combination weights selected by hyper-parameter search and hidden inside the released model. AQI differs from both in exposing the weights: three fixed, positive coefficients that sum to one and can be recomputed from published metric outputs. The interpretability–accuracy trade-off this implies has a direct precedent in MaTESe (Perrella et al., 2022), which traded regression-based opacity for span-level interpretability and, by its authors' own account, matched the top WMT21 submissions on en→de, fell slightly short on zh→en, and dropped sizeably out of domain; our accuracy gap to Ridge and GradientBoosting (§3.5) should be read against that precedent. Judges based on large language models (LLMs), such as GEMBA-MQM (Kocmi and Federmann, 2023), are outside the six-metric inventory audited here (see Limitations).

2 Dataset, Human Ceiling, Metric Audit

2.1 Dataset

All experiments rely on a unified wide-CSV assembled from **two sources**: the bulk comes from the public artefacts of **WMT25 General MT** (Kocmi et al., 2025), and one LP (en→es, $n = 600$) is drawn from the WMT24 release (Kocmi et al., 2024a) as aggregated scores only, kept separate throughout. The corpus has **3,450** segment–system records: a unique translation by one of **40** MT systems for one of **251** unique source segments, over **five** LPs (cs→de, en→cs/es/et/ru). Coverage by the six automatic metrics is 100%; error spans are present on 2,850 segments (82.6%); double independent annotation on **2,391** (69.3%); the corpus was annotated by 34 unique annotators across the four span-annotated LPs.[1] Per-LP composition is in Appendix B, Table 9.

The en→es caveat. These 600 segments use multi-annotator averaging (each segment annotated 1–14 times, mean $\approx$ 4.25), no individual annotator identifiers, and no error spans. We therefore exclude en→es from the inter-annotator statistics in §2.2 and from anything requiring a penalty or pairwise comparison, but use it in §3 for the reference-free scenario.

Domains, systems, protocol. Domains are biased toward *literary* (41.9%) and *news* (40.8%), with en→es unlabelled (17.4%); this limitation is discussed in the Limitations section. The 40 systems span commercial LLMs, open-source LLMs, MT-specialised models and WMT25 submissions. Annotation follows the **ESA** protocol (Kocmi et al., 2024b), an MQM-derived

[1] Annotator pools per LP, from the WMT25 findings (Kocmi et al., 2025, Table 6): cs→de and en→cs—ÚFAL Charles University linguists, researchers and students, native in one language with high proficiency in the other; en→et—professional translators (Luisa Language Solutions); en→ru—Toloka paid expert crowd, bilingual native speakers of the target language. All were professional translators but not trained MQM/ESA annotators, completed an interactive tutorial and a fixed control task used to model annotator reliability, and were paid a fair local wage; each segment was annotated exactly twice.

scheme: the annotator first marks **error spans** with severity $\in$ {minor, major} (no error categories are recorded), then assigns a **holistic score** on 0–100 informed by the marked spans. We aggregate spans into an MQM-style **span penalty** $P_a = |E_a^{\text{minor}}| + 5|E_a^{\text{major}}|$; segment-level `score`/`penalty` are the mean over the two annotators. Tier boundaries follow the industry MQM convention: **Good** ≥ 85, **Medium** $60 \leq s < 85$, **Bad** < 60 (empirical shares 43.0/35.0/22.0%). Span-dependent targets are computed on 2,504 segments (after excluding en→es and NaN rows); the ceiling statistics in §2.2 on the 2,391 double-annotated segments. The slice size is bounded by the cost of running the learned metrics at scale.

2.2 The human ceiling

Any automatic metric calibrated against human assessment is bounded above by the agreement between humans themselves. In this subsection we quantify this ceiling at three levels—overall score, span penalty, and binary error presence—and establish the theoretical upper bound for AQI tier-agreement.

2.2.1 Pairwise score agreement

On the subcorpus of $n = 2{,}391$ double-annotated segments we compute pairwise statistics between two independent scores `score_annotator1` and `score_annotator2`. Pearson correlation $r =$ **0.425**; Spearman $\rho = 0.417$. This is a **moderate** dependence: two annotators, evaluating the same translation under identical instructions, diverge substantially. The distribution of absolute difference $|\Delta| = |s_1 - s_2|$ yields **mean** $|\Delta| = 20.24$ points, **median** $|\Delta| = 15.00$, std $= 18.25$, maximum 97 points (extreme cases are examined below).

Strict agreement (± 5 points) holds in only **22.8%** of segments; agreement within one tier-band—approximately ± 20 points on a scale with tier widths 25/15/60—is reached on **61.3%**. Thus in **38.7%** of cases two annotators disagree by *more than one tier*. This is the empirical upper bound against which we compare AQI tier-match in §3.

Agreement breaks down highly unevenly by LP: `en→et` yields $r = 0.598$ (a bimodal quality distribution—translations are either good or clearly broken), `cs→de` $r = 0.337$, `en→cs` $r = 0.304$, and `en→ru`—$r = 0.113$, essentially noise (per-LP breakdown in Appendix C, Fig. 3). The spread tracks the annotator pool rather than the language (professional translators on en→et, a paid crowd on en→ru; §2.1): the ceiling we report is a property of the annotation setup, not a universal constant. The spread also motivates the per-LP offsets of §3.

2.2.2 Span-penalty agreement and the nature of the noise

Agreement on error-span annotation is systematically *worse* than on the holistic score: on the same subcorpus $n = 2{,}391$, Pearson between the two annotators on the span penalty is $r = 0.279$ (vs. $r = 0.425$ on the holistic score), because span annotation requires three independent decisions per error (presence, location, severity) and noise at each step degrades agreement multiplicatively. This is consistent with prior work rather than new: MQM annotation is known to have low inter-annotator agreement (Freitag et al., 2021), and Kocmi et al. (2024b) report that ESA annotators mark an overlapping span in only $\sim$50% of cases between annotators and $\sim$29% within the same annotator, whereas the final score benefits from being assigned *after*, and primed by, the marked spans. Our contribution is to quantify both levels on the same segments and annotators. Throughout the rest of the paper we use **holistic human score** as ground truth and refer to the span penalty only where explicitly noted. Inspecting the extreme disagreements ($|\Delta s| \geq 84$) shows that they are not data outliers but legitimate borderline cases driven by **stylistic strictness**, **domain-expertise asymmetry**, and **literal vs. creative translation**; three worked case studies are in Appendix D. This is structural noise of the task, which AQI must carry rather than try to resolve.

2.2.3 Implication for AQI

The empirics above imply a hard **theoretical ceiling** on tier-agreement of any automatic metric with the mean-aggregated human annotation. The direct *empirical* estimate: two annotator scores for the same segment agree on tier in only **48.6**% of cases on the canonical $n = 2{,}391$ set (pairs with two *different* annotators; 51.3% on the larger $n = 2{,}850$ set that includes 459 self-agreement pairs—one annotator scoring both slots—used for the human-in-the-loop sweep in §4). A parametric model gives a comparable bound: *assuming* $\Delta = s_2 - s_1 \sim \mathcal{N}(0, \sigma_\Delta^2)$ with $\sigma_\Delta \approx 25$ (from mean $|\Delta| = 20.24 = \sigma_\Delta\sqrt{2/\pi}$; an approximation, since scores are censored at 0 and 100), a

Metric	ρ score	ρ penalty	AUC major
MetricX-XXL	**−0.485**	+0.391	**0.675**
xCOMET-XXL	+0.413	−0.357	0.658
BERTScore F1	+0.406	−0.180	0.599
TER	−0.333	+0.137	0.583
chrF	+0.352	−0.127	0.573
COMETKiwi-XXL	+0.084	−0.062	0.546

Table 2: Per-metric correlations and AUC. Signs follow the scale direction: MetricX-XXL and TER are lower-is-better, the others higher-is-better.

	BERT	chrF	CK	MX	TER	xCOM
BERT	—	+0.86	+0.23	−0.50	−0.80	+0.43
chrF	+0.86	—	+0.17	−0.34	−0.85	+0.29
CK	+0.23	+0.17	—	−0.28	−0.11	+0.44
MX	−0.50	−0.34	−0.28	—	+0.29	−0.70
TER	−0.80	−0.85	−0.11	+0.29	—	−0.26
xCOM	+0.43	+0.29	+0.44	−0.70	−0.26	—

Table 3: Spearman ρ between the six metrics. $n = 3{,}450$. Signs reflect aligned scale directions. **MX** = MetricX-XXL, **CK** = COMETKiwi, **xCOM** = xCOMET-XXL.

simulation over the empirical s_1 distribution yields $\mathbb{P}[\text{tier}(s_1) = \text{tier}(s_2)] \approx \mathbf{0.60}$: even an ideal oracle reproducing one annotator's score would disagree with the paired annotator on tier in roughly 40% of cases. The calibrated F7+LP variant reaches 53.0% tier-match (§3)—slightly above the 48.6% empirical ceiling and within ≈ 2 percentage points (pp) of the $n = 2{,}850$ figure; the narrow excess is the Spearman-Brown effect (§2.3.1): F7+LP is evaluated against the mean-aggregated score, a smoother target than a single annotator. In a task with inter-annotator $r = 0.425$, expecting substantially higher agreement is metrically incorrect.

2.3 Audit of six metrics

All six metrics from §1.2—chrF, TER, BERTScore F1, COMETKiwi-XXL, MetricX-XXL (reference-based) and xCOMET-XXL—are computed on the full corpus of 3,450 segment-system pairs. The audit aims to establish (i) which metrics correlate most strongly with human score, (ii) which of them carry overlapping signal and are therefore redundant in a linear combination, and (iii) which single metric best detects the presence of a major error.

2.3.1 Correlation with human score and AUC for major-error detection

Table 2 gives the segment-level Spearman ρ of the metrics against the mean-aggregated score ($n = 3{,}450$) and AUC for the binary task of detecting "≥ 1 major span" (on 2,504 segments with span annotation).

MetricX-XXL has the strongest signal of the six, with $|\rho| = 0.485$ against the human score. It is tempting to note that this exceeds the inter-annotator $r = 0.425$, but **the two numbers are not directly comparable**: the metric is correlated against the *mean-aggregated* score (two or more annotators averaged), whereas 0.425 is a single annotator against another single annotator. Averaging denoises the target, which mechanically inflates any correlation with it. On a fair, like-for-like basis—MetricX-XXL against a *single* annotator, measured on the $n = 2{,}850$ two-annotator set (§4)—the correlation is $\rho \approx 0.46$, i.e. **at, not above**, the single-vs-single inter-annotator agreement on that set ($\rho = 0.465$; the figure rises to 0.52 only when the metric is scored against the *averaged* target, which is the inflation we are flagging). The honest reading is that the best single metric *approaches* the human ceiling from below; no automatic metric beats it. The worst is **COMETKiwi-XXL** with $\rho = 0.084$, effectively zero signal. The explanation is a **calibration mismatch**: the COMETKiwi model was trained mainly on the news domain of WMT-17–22, whereas 41.9% of our corpus is literary and a substantial share is speech-like dialogue. The calibration does not transfer: COMETKiwi yields a compressed range of scores on out-of-domain segments, and the linear correlation with humans collapses. The AUC for major-detection follows the same order (Table 2), with COMETKiwi (0.546) close to chance.

2.3.2 Pairwise redundancy

The Spearman matrix between the six metrics (Table 3; heatmap in Appendix C, Fig. 4) reveals three nearly collinear pairs within the surface family.

The three strongest within-family connections are **chrF** ↔ **BERTScore F1** ($|\rho| = 0.855$), **chrF** ↔ **TER** ($|\rho| = 0.852$), and **BERTScore F1** ↔ **TER** ($|\rho| = 0.803$). All three metrics measure different projections of the *same reference-overlap signal*: chrF at the level of character n-grams, BERTScore at the level of token embedding similarity, TER at the level of edit distance. Their triplet is nearly triple-redundant: the variance inflation factor (VIF) for any pair from the three exceeds 5, which in linear regression creates multicollinearity and instability of weights. Including all three in a single composite formula is equivalent

to *triple counting* of the same signal at the expense of the weight budget; zeroing two of three (while keeping one surface baseline for interpretability) is the correct solution, which is also what the grid search in §3 converges to.

Between families the picture is different. **MetricX-XXL** $\leftrightarrow$ **xCOMET** yields $|\rho| = 0.699$—a noticeable connection (both are neural metrics), but not collinearity: $\approx 50\%$ of each metric's variance is not explained by the other. **COMETKiwi** falls outside all clusters: its maximum connection is with xCOMET ($\rho = 0.438$), its minimum with TER ($\rho = -0.107$). This is consistent with its low correlation with human score: COMETKiwi carries neither the surface signal (like chrF/BERT/TER) nor a strong semantic one (like MetricX/xCOMET), and on our domain produces a high-variance, low-informative output. The effective number of independent signals in the inventory is approximately **four**: MetricX-XXL, xCOMET, a surface baseline (one representative of chrF/BERT/TER), and the human span penalty.

2.3.3 Implications for building the composite

The audit establishes three facts that motivate the closed form of AQI in §3: (i) no single metric exceeds $|\rho| = 0.5$ against human score, and even the best (MetricX-XXL) only *approaches* the inter-annotator ceiling rather than clearing it (§2.3.1)—a combination is needed to push past a single metric; (ii) three of six metrics (chrF, BERTScore, TER) are nearly triple-redundant, so a composite should keep only *one* representative of the surface cluster rather than all three; (iii) COMETKiwi on multi-domain WMT25 yields almost zero signal due to calibration mismatch and should receive zero weight in any linear combination trained on this corpus.

2.4 Normalising metrics to a common scale

The six metrics live on different scales and have different orientations: chrF is in $[0, 100]$ higher-is-better and TER in $[0, +\infty)$ lower-is-better; BERTScore F1, xCOMET-XXL and COMETKiwi-XXL are in $[0, 1]$ higher-is-better, while MetricX-XXL is in $[0, 25]$ *lower-is-better* (it is an error score, not a quality score). To combine them in a linear sum we map each to a unified $[0, 100]$ scale, higher-is-better. For five of the six the normalisation is trivial rescaling (and, for TER, inversion); the exact forms are in Appendix J.

For **MetricX-XXL** the scale is an error count and "quality" does not move linearly from 0 to 25 (the gap between 1 and 5 feels sharper than between 15 and 19). We adopt the linear normalisation

$$\text{mx_lin} = \text{clip}\big(1 - \tfrac{\text{MetricX-XXL}}{25},\ 0,\ 1\big) \cdot 100,$$

which is scale-preserving (each error unit lowers the score by 4 points), tier-friendly (its distribution covers $[16, 99]$, aligned with the 60/85 boundaries), and free of any hyperparameter. As a single-metric sanity check, mx_lin is an affine transform of raw MetricX-XXL, so it preserves both the Spearman rank correlation (0.4851) and the Pearson correlation (0.4500) of the raw metric against the human score. We also tried a non-linear exponential normalisation and other parametric forms; none gave a robust cross-validated gain over mx_lin, and the exponential variant is carried as an equal-billing alternative in Appendix E.

3 The AQI Formula: Derivation and Justification

3.1 Problem statement and normalisation

This section derives the **auto-only** part of AQI—a formula computed from automatic metrics alone. §4 then shows how an optional human score is folded on top; here we restrict the discussion to the purely automatic case. The formula must produce a scalar in $[0, 100]$ that increases monotonically with the human score and is transparent enough to be reproduced and interpreted without a black box.

Normalisation of the six metrics to the unified $[0, 100]$ higher-is-better scale is introduced in §2.4; for MetricX-XXL we use the linear form $\text{mx_lin} = (1 - \text{MetricX-XXL}/25) \cdot 100$ throughout the main text (an exponential alternative is in Appendix E).

All experiments in this section use the unified corpus of §2.1 ($n = 3{,}450$; mostly WMT25 General MT plus the separate en→es pair; the CSV ships with the repository). Tier partition is the paper-wide standard: **Good** ≥ 85, **Medium** $60 \leq s < 85$, **Bad** $s < 60$. All fitted models use 5-fold cross-validation (CV; `KFold`, `shuffle=True, random_state=42`); reported correlation and mean absolute error (MAE) are on out-of-fold predictions (no leakage).

3.2 Single metrics as a baseline

The first reference point is the standalone performance of each metric. This gives a lower bound for any composite: a linear combination has to

Metric	Pearson	Spear.	MAE	Tier
mx_lin	**+0.450**	**+0.485**	**17.61**	**46.4%**
xc100	+0.442	+0.481	20.08	48.2%
bert100	+0.374	+0.413	18.06	44.8%
chrf100	+0.289	+0.311	24.92	30.6%
ter_inv100	+0.023	+0.300	31.41	29.0%
ck100	+0.138	+0.163	21.60	38.1%

Table 4: Single metrics, $n = 3{,}450$, tier 85/60. MetricX-XXL with linear normalisation is the best by Pearson. COMETKiwi-XXL yields a weak correlation due to a domain mismatch (§2.3).

justify itself against the best single metric—if it does not improve on it, it is redundant. The bar here is MetricX-linear (Pearson 0.450, Spearman 0.485); we return to whether our composite clears it in §3.5.

3.3 Candidate metric set: MetricX-XXL + xCOMET + BERTScore

Before tuning weights we have to decide which metrics to combine, and the audit in §2.3 answers this directly. Only three metrics give a **strong univariate signal** against the human score—MetricX-XXL, xCOMET and BERTScore F1 (Table 4, $\rho \in [0.37, 0.49]$); chrF and TER are noticeably weaker ($\rho \in [0.02, 0.29]$), and COMETKiwi degrades to $\rho \approx 0.14$ on our multi-domain corpus. The surface trio chrF, BERTScore F1 and inverted TER is nearly triply collinear ($|\rho| \geq 0.80$ pairwise), so we keep a **single representative**: BERTScore F1, which has the highest univariate correlation within the cluster (0.406 vs. chrF 0.352 vs. TER 0.023 on Pearson) and whose embedding similarity adds a **semantic** component that character- and edit-level overlap cannot. The candidate set is therefore *MetricX-XXL + xCOMET + BERTScore F1*: the two strongest learned metrics (with $\approx 50\%$ non-shared variance) plus one surface-cluster representative.

3.4 Families of formulas we tried

3.4.1 Weighted linear sum (the F family)

The simplest composite is a linear combination $\text{AQI} = \sum_i w_i \cdot m_i$ with weights $w_i \in [0, 1]$, $\sum_i w_i = 1$. The constraints have a natural reading: "share of each metric in the final score". We searched 3-metric combinations on the candidate triple on a grid of step 0.05. The best point on the joint balance of Pearson, tier-agreement, and MAE—we call it F7:

$$\text{AQI}_{F7} = 0.40 \cdot \text{mx} + 0.30 \cdot \text{xc100} + 0.30 \cdot \text{bert100} \quad (1)$$

where *mx* is the normalised MetricX-XXL (mx_lin). The 40%/30%/30% weights are confirmed by a constrained least-squares fit with Sequential Least Squares Programming (SLSQP; $\sum w_i = 1$, $w_i \geq 0$) on the full data: $(0.396, 0.287, 0.317)$ for linear MetricX—essentially the manual rounding. For transparency we fix the rounded $(0.40, 0.30, 0.30)$.

3.4.2 Geometric mean (the Geo family)

A geometric mean multiplicatively penalises single-metric failures. Geo3 and Geo4 are competitive with F7 on Pearson (0.49) but, without calibration, collapse to 28% tier-agreement because of a compressed distribution, and adding TER and chrF (Geo5) breaks down to $\rho = 0.18$; full results are in Appendix F, Table 11.

3.4.3 Linear regression on all six metrics (Ridge)

An unconstrained Ridge regression ($\alpha = 1$) on all six standardised metrics reaches $\rho = 0.517$ (no LP), $\rho = 0.612$ (+LP)—higher than F7-linear, but with one odd detail: the COMETKiwi-XXL coefficient is **negative** (-0.20) despite its positive univariate correlation—the *suppressor effect* (Conger, 1974), since the metric is near-zero signal (§2.3.3). Dropping it, **Ridge-5 (no CK)** has all-positive coefficients at almost no accuracy cost ($\rho = 0.491$ / 0.62 with LP). So the suppressor is not a reason to reject linear regression; we still publish F7 as the headline for *parsimony*—three metrics instead of five, nearly tying Ridge-5 on Pearson (0.483 vs. 0.491). The full Ridge coefficients are in Appendix E.

3.4.4 Non-linear tabular machine-learning (ML) models

To bound the **ceiling** of any auto-only regressor on the same six metrics we ran standard tabular-ML models (Table 5); these are not finalists (no closed form). Neural regressors on raw src/ref/tgt text (Appendix H) do not fit the "formula on metric outputs" requirement either; the best ($\rho \approx 0.80$) is disqualified by label leakage.

3.5 Master comparison

F7 clears the redundancy bar of §3.2: without calibration it reaches Pearson 0.483 against the

Model	Pearson	Spear.	MAE	Tier
GradientBoosting	+0.571	+0.559	13.63	51.4%
RandomForest ($n = 200$)	+0.558	+0.546	13.80	51.3%
KNN ($k = 15$)	+0.549	+0.541	13.81	51.3%
MLP (32×32)	+0.544	+0.539	13.98	49.9%

Table 5: Non-linear ML models on the same six metrics, without human error spans (KNN: k-nearest neighbours; MLP: multilayer perceptron). The best is GradientBoosting ($\rho \approx 0.57$)—the auto-only regression ceiling on this feature set without LP calibration. Not adopted as the primary formula: see the interpretability trade-off in §3.5.

Formula	Pearson	MAE	Tier
F7-linear	**+0.483**	**14.55**	**46.5%**
F7-linear + LP	+0.624	13.08	53.0%
AQI-Geo3	+0.492	28.32	28.3%
AQI-Geo3 + LP	+0.579	14.65	54.8%
Ridge (6 metrics)	+0.517	14.31	47.9%
Ridge-5 (no CK)	+0.491	14.64	45.5%
Ridge + LP	+0.612	13.20	53.1%
GradientBoosting	+0.571	13.63	51.4%
xCOMET alone	+0.442	20.08	48.2%
MetricX-linear alone	+0.450	17.61	46.4%

Table 6: Master comparison on the full corpus $n = 3{,}450$, 5-fold CV, tier 85/60. All variants are auto-only (no human error spans, no human score at inference).

best single metric's 0.450—a modest but bootstrap-reliable gain ($\Delta\rho = +0.033$, 95% confidence interval $[+0.016, +0.050]$)—and separates the low-quality tail better (0.341 vs. 0.306 on segments scoring < 60). It does *not* win on raw accuracy: Ridge (0.517), Ridge-5 (0.491) and the GradientBoosting ceiling (0.571) are higher. The honest framing is a trade-off: F7 gives up ≈ 0.03 Pearson against Ridge and ≈ 0.09 against GradientBoosting for a one-line formula with three positive weights. We *hypothesise* that for a product-facing index used by project managers and clients this closed form is worth the gap; we do not test the preference with stakeholders directly, but ground it in three properties the black-box alternatives lack. (i) AQI is monotone in every input by construction, whereas GradientBoosting lowers its score on 4.0% of out-of-fold segment pairs in which all three inputs strictly improve (median drop 3.1 points, 90th percentile 12.4). (ii) Three published weights are a versionable specification that can be written into a client or vendor agreement and recomputed in a spreadsheet; a fitted model with post-hoc attribution explains individual predictions but does not keep scores comparable across retrains, and attribution over the near-collinear surface cluster (§2.3) is unstable by construction. (iii) With per-LP offsets the gap closes: F7+LP reaches 0.624 against Ridge+LP 0.612 and GradientBoosting+LP ≈ 0.62 under the same offset protocol. A team needing only a ranking signal should still prefer GradientBoosting or Ridge-5. Geo variants match F7+LP on tier only with calibration; without offsets they lose $\sim$18 pp. We do not claim F7 is more robust *across* LPs—its per-LP spread is no tighter than MetricX-linear's; its advantage is on the aggregate, ranking, and the low-quality tail.

The gain is asymmetric. On the 5 % of segments where F7 and MetricX-alone disagree most ($|\Delta| \geq 8.5$ points after linear calibration of both to the human scale, $n = 173$), F7 is closer to the human score in 57 % of cases; when the composite *raises* a segment that MetricX penalises it is right in 69 % of cases ($n = 74$), when it *lowers* one in 49 % ($n = 99$). Disagreements split evenly between segments with and without a major span; the composite's edge is weakest on news (F7 closer in 45 % of news disagreements vs. 55 % on literary), the domain where MetricX alone is strongest (§3.7). Two worked examples are in Appendix D, Cases 4–5.

3.6 F7 as the final formula

We pick F7 as the final auto-only AQI: closed-form, three positive weights, no free parameters. With the linear MetricX-XXL normalisation it gives a tier-friendly scale that covers $[0, 100]$ roughly evenly without any calibration. The final formula:

$$\begin{aligned}\mathrm{AQI}_{\mathrm{linear}} = {} & 0.40 \cdot \left(1 - \tfrac{\mathrm{MetricX}}{25}\right) \cdot 100 \\ & + 0.30 \cdot \mathrm{xCOMET} \cdot 100 \\ & + 0.30 \cdot \mathrm{BERTScore} \cdot 100 \qquad (2)\end{aligned}$$

An equal-billing exponential variant, identical in ranking and Pearson but on a compressed scale, is given in Appendix E.

3.7 Language-pair (LP) offsets—optional calibration

The tables above repeatedly mention "+ LP". Defined once: a **language pair (LP) offset** δ_{LP} is a constant added to AQI on segments of a given LP to compensate for the systematic per-LP bias we observe in F7 residuals. The offset is computed on

LP	n	Offset
cs→de	600	-0.98
en→cs	1,050	$+2.11$
en→es	600	$+10.29$
en→et	600	-16.75
en→ru	600	-1.83

Table 7: Per-LP residual offsets for F7-linear $(0.40, 0.30, 0.30)$. The mixed signs reflect per-LP difficulty: en→et is over-rated by the metrics by $\sim$17 points, en→es is under-rated by $\sim$10.

labelled data (human scores are needed for calibration, but not at inference) and only applies to **previously seen LPs**—for a new LP not in the training corpus, no offset is available and the formula runs without it. We publish offsets as an **optional extension**, not as part of the core formula.

Table 7 gives the mean of score $-$ AQI$_{F7}$ (5-fold CV) by LP.

Calibrated version: AQI$^{+\text{LP}}$ = AQI + δ_{LP}, where δ_{LP} comes from Table 7. On 5-fold CV (offsets estimated on the train fold, applied on the test fold) they lift F7-linear from Pearson 0.483 / tier 46.5% / MAE 14.55 to **0.624** / **53.0**% / **13.08** (Table 6): $+0.141$ Pearson and $+6.5$ pp tier-agreement.

Default recommendation: use the **core formula without offsets**. Offset calibration is for those working with the same LPs as the canonical WMT25 corpus, who want the additional $+0.12$–0.14 Pearson and $\sim$6–7 pp tier-agreement.

Domain dependence. The corpus has two labelled domains, *literary* (n = 1,444) and *news* (n = 1,406), confounded with LP (cs→de is news only). Table 12 (Appendix K) reports F7, MetricX-alone and an in-sample per-domain refit of the three weights. The optimum does move: literary favours $(0.60, 0.05, 0.35)$, news collapses almost to MetricX alone $(0.90, 0.10, 0.00)$; the refit gains $+0.03$–0.04 Pearson within a domain, where F7 trails MetricX-alone by 0.01–0.03, and weights fitted on one domain transfer to the other within 0.02 of the local optimum. F7's advantage is on the mixed corpus it was fitted on and on the low-quality tail (0.338 vs. 0.297 on literary segments scoring < 60). We therefore recommend the fixed weights for mixed production traffic and, for a single known domain, the recalibration recipe in the Limitations section.

4 Human-in-the-Loop

4.1 Why human-in-the-loop

Auto-only F7-linear reaches tier-agreement 46.5% (§3.5), *below* the independent inter-annotator tier-agreement of 48.6% (§2.2; 51.3%, with $r = 0.467$, on the n = 2,850 two-score set used for the α sweep below, which includes 459 self-agreement pairs)—it approaches human consistency from below. "Human-calibrated" means *trained against* human scores, not human parity. For low-stakes use (ranking, post-editing prioritisation) auto-only AQI suffices, but in high-stakes scenarios (legal, medical, brand-critical) undetected disagreements carry non-linear risk and human review is required. We therefore derive a **single hybrid index** blending one human score into the auto-only AQI and select its weight empirically. Since AQI is normalised to the same $[0, 100]$ higher-is-better scale as the human score (§2.4), $(1-\alpha)\cdot$AQI$+\alpha\cdot$human is a well-defined convex combination.

4.2 The hybrid index

A segment is usually scored by one annotator (the base case); with several, we average their scores. The hybrid index is

$$\text{AQI}_H = (1-\alpha)\cdot\text{AQI}_{F7} + \alpha\cdot\text{human} \quad (3)$$

where human is the single annotator's score (or mean) on the native $[0, 100]$ scale. The human score enters **as-is**: no shrinkage toward a prior, no per-annotator z-normalisation. A human scoring a segment low may be correctly catching a defect the index missed, so pulling them toward a prior would treat the index as the arbiter of truth—the opposite of why we bring in a human. Dropping z-normalisation is also practical: there is rarely enough data per annotator to estimate μ_i, σ_i reliably.

4.3 Choosing α

Tuning α on the same human score that enters the formula would be circular. The validation is therefore **honest, on an independent annotator**: we blend in annotator-1's score and evaluate the result against the held-out annotator-2 score (which the formula never saw), then symmetrically blend in annotator-2 and evaluate against annotator-1, and average the two directions. The α sweep runs on the $n = 2850$ segments with two annotator scores (see §2.2 for the note on self-agreement pairs).

α	Pearson	Tier	MAE
0.00 (auto-only)	0.425	40.1%	17.47
0.20	0.520	44.7%	16.48
0.25	0.527	46.3%	16.34
0.30 (final)	**0.529**	47.1%	16.24
0.35	0.528	48.0%	16.17
0.40 ($= w_{\text{mx}}$)	0.525	48.9%	16.15
0.50	0.517	50.5%	16.23
1.00 (pure human)	0.467	51.3%	18.80

Table 8: Independent-annotator α sweep ($n = 2850$). The selection criterion is Pearson ρ against the independent annotator: tier-agreement rises with α trivially, because at high α the prediction simply copies the blended human's tier labels, whereas correlation against an *independent* human is the honest signal.

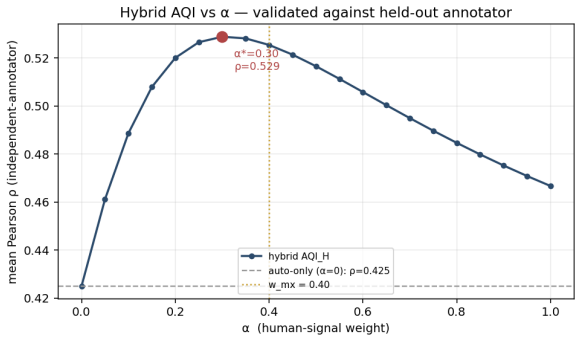

Figure 1: Mean Pearson ρ (independent-annotator validation) as a function of α. The curve rises from auto-only ($\alpha = 0$, $\rho = 0.425$) to a flat optimum around $\alpha \approx 0.25$–0.40 peaking at $\alpha = 0.30$ ($\rho = 0.529$), then declines gently toward pure-human ($\rho = 0.467$).

The optimum is a **broad plateau** $\alpha \approx 0.25$–0.40 with argmax at $\alpha = 0.30$. The final $\alpha = 0.30$ lifts correlation from $\rho = 0.425$ to $\rho = 0.529$ ($+0.104$, $\approx$24% relative gain over auto-only). Notably, the plateau covers the MetricX-XXL weight in F7 ($w_{\text{mx}} = 0.40$): the human earns roughly the same trust as the strongest single metric—$\alpha = 0.40$ gives $\rho = 0.525$, statistically indistinguishable from the peak. With $\alpha = 0.30$, Eq. 3 becomes the flat four-term formula given in §5, which is the final human-in-the-loop index.

5 Conclusion

We have presented the **Aggregate Quality Index (AQI)**—a composite, human-calibrated index of machine translation quality, developed on a unified corpus of $n = 3{,}450$ segment-system ratings, drawn primarily from WMT25 General-MT (en→es from WMT24, see §2.1). The work makes **four contributions**: (i) an empirical quantification of the inter-annotator agreement ceiling $r = 0.425$ on the double-annotated $n = 2{,}391$ subset—the theoretical upper bound for any automatic metric; (ii) a systematic audit of six metrics with per-pair redundancy analysis, identifying the COMETKiwi-XXL domain mismatch and the near-triple collinearity of chrF / BERT / TER; (iii) a closed-form formula F7-linear, reaching $\rho \approx 0.48$ on the auto-only baseline without human error spans and without human input at inference; (iv) a simple human-in-the-loop index—a linear blend of one human score into the auto-only AQI with the final weight $\alpha = 0.30$ selected empirically via honest validation against an independent annotator (Pearson rises from $\rho \approx 0.43$ to $\rho \approx 0.53$).

5.1 The final AQI formulas

Recommended deployment: F7-linear (Eq. 2), $0.40 \cdot \text{mx_lin} + 0.30 \cdot \text{xc100} + 0.30 \cdot \text{bert100}$: Pearson $\rho = \mathbf{0.483}$, tier-agreement **46.5**%, MAE **14.55** on 5-fold CV; three positive weights summing to one, no negative coefficients. Optional LP offsets (Table 7) for previously seen LPs lift it to $\rho \approx 0.62$ and tier 53%.

With one human score (human-in-the-loop, §4). Folding in a single human score with weight $\alpha = 0.30$ and multiplying the 0.70 blend weight through the F7 coefficients yields a single flat formula in four terms:

$$\begin{aligned}\text{AQI}_H = {} & 0.28 \cdot \left(1 - \tfrac{\text{MetricX}}{25}\right) \cdot 100 \\ & + 0.21 \cdot \text{xCOMET} \cdot 100 \\ & + 0.21 \cdot \text{BERTScore} \cdot 100 \\ & + 0.30 \cdot \text{human}\end{aligned}$$

where human is the single annotator's 0–100 score (or the mean across annotators). The four weights sum to one. On independent-annotator validation this lifts correlation from $\rho \approx 0.43$ (auto-only) to $\rho \approx 0.53$ (§4).

5.2 Interpreting the AQI scale

A production user reads an AQI number directly as a quality verdict, so we publish an explicit interpretation scale (Appendix I, Table 13). The thresholds are *not* the human-score tier boundaries 60/85: the F7-linear distribution is **compressed** relative to the human score (on $n = 3{,}450$, mean $= 76.5$, std $= 10$ vs. human mean $= 75.5$, std $= 21$), so applying 60/85 directly would label 76.6% of segments "Medium" and only 5% "Bad"—uninformative. We instead choose breakpoints **60 / 68 / 75 / 83** so that the mean human score of

the segments in each band ($\{53, 61, 70, 79, 88\}$ from Poor to Good) matches the band's label. One caveat: the headline 46.5% tier-agreement is measured on the **three-tier** human partition ($85/60$), not on this five-band deployment scale.

Impact and reproducibility. The weights are fixed here and LP offsets ship as an optional extension; the canonical CSV, the test sets and the training/evaluation scripts are released at `https://github.com/asyaasinkson/aqi_paper_repo`.

Limitations

LP offsets. The optional per-LP intercepts δ_{LP} (§3.7) are estimated only on the train fold and applied to the test fold within 5-fold CV (no leakage), a standard nuisance control; the primary formula is the no-LP variant, and we publish within-LP correlations as a floor invariant under an LP-mean shift.

Generalisation. Our scope is 5 WMT25 LPs. Not studied: low-resource LPs (sparse training data, weak tokeniser coverage), non-Latin scripts beyond Chinese, Japanese and Korean (CJK)—Arabic, Hebrew, Thai, Indic—and right-to-left/vertical scripts. Before applying AQI on a new LP, run a small-scale calibration on $100+$ doubly-annotated segments and check the within-LP metric–human correlation; if $\rho < 0.20$ the formula does not transfer.

Domains. Our slice is 42 % literary and 41 % news; WMT25's speech, social and dialogue domains are absent, and domain is confounded with LP. Per-domain refits shift the weights (§3.7, Table 12). To adapt AQI to a single known domain: collect $100+$ doubly annotated segments, check the within-domain metric–human correlation, estimate a domain offset exactly as the LP offset ($\delta_{\text{dom}} = \overline{\text{score} - \text{AQI}}$), optionally re-grid the three weights on the simplex, and recalibrate the band breakpoints of Table 13.

Metrics get out of date. AQI depends on MetricX-XXL, xCOMET and BERTScore F1, each updated on its own cycle, with distribution shifts that break the normalisation constant (25). Mitigation: periodic recalibration on a static held-out corpus and an open release of updated coefficients.

LLM-based metrics. GEMBA-MQM and similar LLM judges (Kocmi and Federmann, 2023) are outside our six-metric inventory. We excluded them for the reasons that motivate fixed weights in the first place: per-call cost at production volume, non-determinism, and model versioning that would silently move the composite. WMT25 itself used GEMBA-ESA inside its automatic ranking (Kocmi et al., 2025), so integration is a natural next step; how the weights shift when an LLM judge is added is left to future work.

Reference availability and AI bias. Our formula is reference-based; true zero-shot deployment needs QE variants, left to future work. Annotators may also under-rate machine-translated output (Fischer and Läubli, 2020; Hendy et al., 2023); we did not measure this (WMT25 ran no controlled with/without-label experiment), but the human-in-the-loop blend (§4) partially averages it out by retaining 70 % of the automatic signal. Future directions are listed in Appendix L.

References

Anthony J. Conger. 1974. A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement*, 34(1):35–46.

Lukas Fischer and Samuel Läubli. 2020. What's the difference between professional human and machine translation? a blind multi-language study on domain-specific MT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, pages 215–224.

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. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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 WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 578–628.

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.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Mar-

tins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *Computing Research Repository*, arXiv:2302.09210.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 756–767.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation (WMT)*, pages 478–494.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

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

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 (ACL)*, pages 4984–4997.

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 (ACL)*, pages 311–318.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, 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, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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 (WMT)*, pages 841–848.

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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231.

Michal Štefánik, Vít Novotný, and Petr Sojka. 2021. Regressive ensemble for machine translation quality evaluation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1041–1048. Association for Computational Linguistics.

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 (ICLR)*.

A Per-metric descriptions

This appendix expands the six-metric inventory of §1.2 with the scales, training signals and known failure modes referenced in the main text.

chrF (Popović, 2015) computes an F-score over character n-grams of order 1–6 between the hypothesis and the reference, with a parameter β weighting recall relative to precision (typically $\beta = 2$). Unlike BLEU, chrF operates at the character level without tokenisation, which makes it robust on morphologically rich languages (Slavic, Finno-Ugric) and on compounds. On WMT meta-evaluations chrF consistently outperforms BLEU in correlation with human judgement and serves as a standard surface-level baseline. Weaknesses: sensitivity to punctuation and whitespace, no semantic modelling (antonyms and paraphrases are not adequately penalised or rewarded), and the systematic penalisation of creative non-literal translations. The metric is **reference-based**.

TER (Translation Edit Rate; Snover et al., 2006) is defined as the minimum number of edits—insertions, deletions, substitutions and shifts of word blocks—normalised by the reference length, that are needed to transform the hypothesis into the reference. TER is interpretable (it operationalises post-editing effort), but it operates at the level of exact word matches and therefore systematically penalises synonym replacement, word-order rearrangement and legitimate paraphrasing. The authors proposed human-targeted TER (HTER, with a post-edit as a pseudo-reference) to mitigate this, but in its plain form TER remains a **reference-based** metric with well-known sensitivity to stylistic freedom in translation. In our audit TER serves as a contrast to chrF: both are surface-level metrics, but their optimisation directions are opposed (lower-is-better versus higher-is-better).

BERTScore F1 (Zhang et al., 2020) estimates semantic similarity through greedy token matching between the candidate and the reference based on the cosine similarity of contextual embeddings from a pre-trained BERT/XLM-R model, aggregated into precision, recall and F1. The metric circumvents the lexical rigidity of chrF/TER, correctly crediting synonyms and paraphrases, and correlates well with human judgement on semantic tasks. Known failure modes: the *antonymy problem*—antonyms lie close in the embedding space, so substituting "best" with "worst" yields a high F1; insensitivity to factual inversions of numbers, dates and named entities (e.g. "2 %" and "20 %" yield $F1 \approx 0.97$); and dependence on the training frequency of tokens. **Reference-based**.

COMETKiwi-XXL (Rei et al., 2022b, 2023) is a **reference-free** (QE) regression encoder built on XLM-R-XXL (≈ 10.5 B parameters), taking a source–hypothesis pair and predicting the human assessment. It was trained on Direct Assessments from WMT-17 through 22, MLQE-PE and MQM signals; in the WMT-22/23 QE Shared Tasks it placed among the top systems for both high- and low-resource languages. Strengths: it does not require a reference (applicable in production zero-shot) and covers more than 100 languages. Weaknesses: sensitivity to domain shift (literary, speech), bias toward fluency over adequacy, and a tendency to over-rate "smooth" hallucinations that lack a semantic anchor in the source.

MetricX-XXL (reference-based) is Google's submission to the WMT-23 Metrics Shared Task (Juraska et al., 2023): a regression encoder built on mT5-XXL, fine-tuned in two stages—first on Direct Assessments, then on MQM ratings, with the inclusion of synthetic perturbations. It is released in two families of variants: **reference-based** (`google/metricx-23-xxl-v2p0`, accepting hypothesis + reference) and **QE** (`google/metricx-23-qe-xxl-v2p0`, accepting source + hypothesis). We use the **reference-based XXL variant**. The output is an error score on a $[0, 25]$ scale, lower-is-better. On WMT-23/24 MetricX-XXL occupies top positions in segment-level Spearman correlation with MQM. Weaknesses: it penalises creative and stylistically free transla-

tions, the training signal is biased toward literal correctness, and the metric is sensitive to domain mismatch between training and application.

xCOMET-XXL (Guerreiro et al., 2024) is a hybrid encoder on XLM-R-XXL that combines sentence-level regression with word-level error span tagging following the MQM severity scheme (minor / major / critical). It was trained on an ensemble of Direct Assessments, post-edits and MQM. It supports both modes: **reference-based** (src + hyp + ref) and **reference-free**; we use the reference-based mode. On WMT-23 xCOMET reached state-of-the-art results simultaneously at sentence-, system- and error-span levels and is reported to be the most robust to hallucinations and critical errors. Weaknesses: span boundaries are noisy on short segments, the severity classifier compresses different error types (terminology vs. adequacy) into a single class, and the metric remains vulnerable to adversarial fluent outputs that lack a source anchor.

B Dataset composition

Table 9 gives the full per-LP composition of the 3,450-segment slice referenced in §2.1.

C Inter-annotator agreement and metric redundancy: figures

Figure 2 shows the raw scatter of the two annotators' scores on the $n = 2{,}391$ ceiling set (§2.2); Figure 3 breaks the inter-annotator Pearson r down by language pair; Figure 4 is the heatmap of the pairwise metric Spearman matrix (Table 3).

D Case studies: extreme disagreements

To illustrate the nature of the noise we selected from the 2,391 double-annotated segments three cases with an extreme gap **in the holistic human score** between the two annotators ($|\Delta s| \geq 84$ points across different LPs). These are not data outliers but legitimate borderline cases where annotators saw the same translation differently. Each case below quotes the real source, system output and reference (disputed fragment in **bold**) and analyses a qualitatively different *kind* of disagreement.

Case 1—en→ru, literary, system Algharb. Score: $s_1 = 0$, $s_2 = 97$, $|\Delta| = 97$. MetricX-XXL $= 4.34$, xCOMET $= 0.577$, chrF $= 41.7$.

- **Source:** …"Look at that. You didn't even drop an F-bomb this time." …

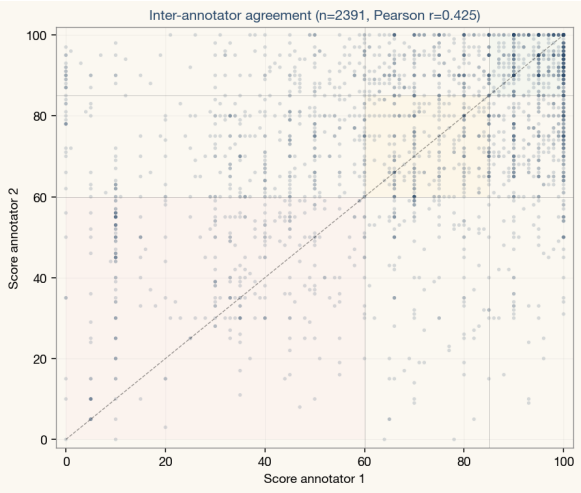

Figure 2: Scatter of scores from two independent annotators on $n = 2{,}391$ segments. The green zone denotes Good × Good, yellow Medium × Medium, red Bad × Bad. The characteristic "cloud" with density along the diagonal but with a substantial fraction of off-diagonal points is visible: in 38.7 % of cases the annotators disagree by more than one tier-band.

- **System:** …"Смотри-ка. В этот раз ты даже не выругался."
  (translit.) "Smotri-ka. V etot raz ty dazhe ne vyrugalsya."
  (gloss) "Look at that. You didn't even swear this time." (English-style straight quotes)
- **Reference:** …— Смотри-ка. В этот раз без лишних эпитетов.
  (translit.) — Smotri-ka. V etot raz bez lishnih epitetov.
  (gloss) "— Look at that. This time without unnecessary epithets." (native Russian em-dash)

Annotator 1 marked eight minor spans (mostly zero-length—punctuation, quotes, the choice of quotation styles) and gave a score of **0**. Annotator 2 found no errors and gave **97**. The translation is essentially correct and reasonably fluent; the disagreement is fully explained by **threshold strictness toward style**: one annotator perceives the dialogue formatting via straight quotes instead of the Russian em-dash as a series of microerrors, the other as acceptable variation. This class of disagreement shows that a score of 0 does not always mean "bad translation"—it can mean "strict annotator on a stylistically neutral segment".

Case 2—cs→de, news, system CommandA. Score: $s_1 = 7$, $s_2 = 95$, $|\Delta| = 88$. MetricX-XXL

LP	seg.-sys.	unique segs.	systems	annot.	mean	std	spans
cs→de_DE	600	30	21	10	79.04	18.58	yes
en→cs_CZ	1,050	53	20	11	76.86	15.97	yes
en→es	600	102	14	1–14†	90.67	14.37	no
en→et_EE	600	33	19	8	54.72	21.87	yes
en→ru_RU	600	33	19	5	75.25	20.08	yes
Total	**3,450**	**251**	**40**	**34**	**75.51**	**21.08**	**2,850**

Table 9: Dataset composition by language pair. The unique-source-text counts reflect our 3,450-segment slice after filtering and stratification; the full WMT25 corpus contains substantially more unique segments. The "annot." column gives the number of unique individual annotators per LP. † en→es uses a different annotation protocol: 1–14 annotations per segment (mean ≈ 4.25, median $= 1$; 153 segments with 10 annotations, 45 with 8, 20 with 5 etc.), aggregated into a single mean score without preserving individual annotator identifiers.

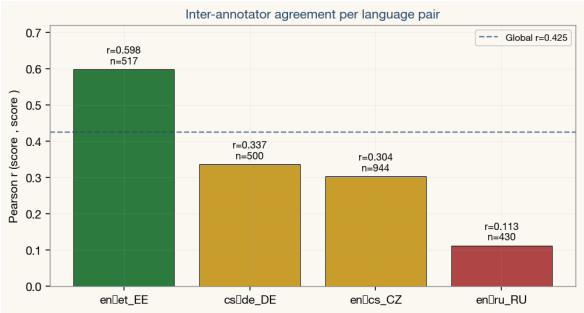

Figure 3: Inter-annotator Pearson r per language pair. Green bars—$r \geq 0.4$ (acceptable agreement), yellow—borderline, red—$r < 0.25$ (essentially noise). The dashed line marks the global $r = 0.425$. The five-fold gap between en→et ($r = 0.598$) and en→ru ($r = 0.113$) reflects different semantic difficulty across LPs and annotator subjectivity.

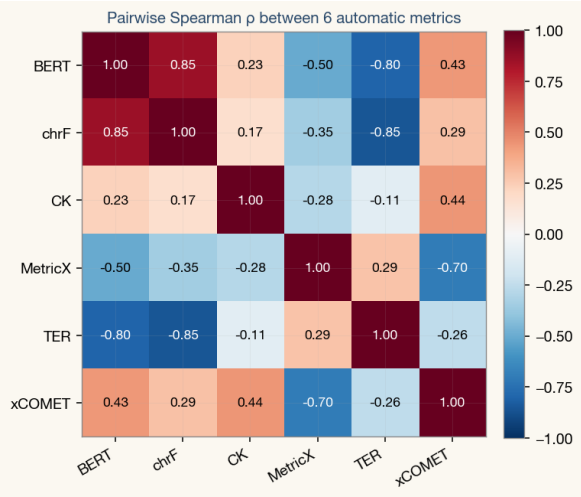

Figure 4: Pairwise Spearman ρ between the six automatic metrics on $n = 3{,}450$. Three collinear clusters are visible: the surface family (BERT/chrF/TER) with $|\rho| \geq 0.80$; the neural family (MetricX/xCOMET) with $|\rho| = 0.70$; and the outlier COMETKiwi with near-zero connections to the surface family.

$= 5.69$, xCOMET $= 0.465$, chrF $= 62.2$.

- **Source:** …vázaných na horninu zvanou **hadec** jako **kuřička Smejkalova** nebo houba zvaná **kuřátka něžná** …
- **System:** …an den **Schlangenstein** gebundener Arten wie der **Smejkal-Schlangenkornblume** oder des Pilzes **Kuřátka něžná** …
- **Reference:** …an das Gestein **Serpentinit** gebunden sind – wie **Mieren** oder der Pilz **Zierliche Koralle** …

The translation contains four terminology errors at once: *hadec* (a mineralogy term, "serpentinite") is rendered as *Schlangenstein* (a literal calque); *kuřička Smejkalova* (a plant species, "Smejkal's mouse-ear chickweed") is rendered as the non-existent *Smejkal-Schlangenkornblume*; and *kuřátka něžná* (a mushroom species) is left untranslated. Annotator 1 marked four major spans and one minor and gave **7**, capturing a severe terminology failure. Annotator 2 marked one minor (the same terminology error by localisation) and gave **95**, evidently lacking the domain knowledge to evaluate mineralogical and botanical accuracy. MetricX-XXL $= 5.69$ (above our has-major threshold $= 4.79$) correctly signals the presence of a major error—that is, the automatic metric is *more accurate* in recognising the defect than the second annotator. The case illustrates **domain-expertise asymmetry** as a source of noise.

Case 3—en→cs, literary, system GemTrans. Score: $s_1 = 93$, $s_2 = 9$, $|\Delta| = 84$. MetricX-XXL $= 2.62$, xCOMET $= 0.811$, chrF $= 49.2$.

- **Source:** Kyle cleared his throat and **leaned back**, his expression softening …

- **System:** Kyle si odkašlal a **opřel se zpět**, jeho výraz se zmírnil … (literal)
- **Reference:** Kyle si odkašlal a **zhoupnul se v židli**. Jeho výraz se uvolnil … (idiomatic)

Annotator 1 found no errors and gave **93**; both MetricX-XXL (2.62) and xCOMET (0.811) agree—the translation is good. Annotator 2 marked six spans (two major, four minor) and gave **9**. Analysis of the spans shows that this annotator penalises **literal vs. creative translation**: "leaned back" rendered as a normative literal translation is judged as major; paraphrasing instead of literal transfer of conversational interjections is judged as minor. This class of disagreement is **stylistic purism of a single annotator** not backed by an objective translation defect—here it is not the metric that penalises creative translation but the human, diverging from both automatic metrics and the second annotator.

The three cases represent three qualitatively different sources of disagreement—stylistic strictness, domain-expertise asymmetry, and literal vs. creative translation—and confirm that a substantial share of disagreements does not reduce to "one is right, the other is wrong".

Where the composite and MetricX disagree. The two cases below come from the disagreement set of §3.5: the top 5 % of segments by $|\text{F7} - \text{MetricX}|$ after linear calibration of both to the human scale. Case 4 is the frequent pattern (the composite *raises* a segment MetricX punishes and is right); Case 5 the coin-flip pattern (the composite *lowers* one and is wrong).

Case 4—en→ru, news, system Claude-4. Human 93.5 ($s_1 = 93$, $s_2 = 94$; one minor span each). MetricX-XXL = 15.1 (top 1 % of the corpus: only 30 of 3,450 segments score higher), so mx_lin = 39.5; xCOMET = 0.41 (bottom 5 %); BERTScore = 0.78 (near the corpus median). F7 = 51.4 vs. MetricX-alone 39.5.

- **Source:** …forming a crucial **second-wicket stand** of 93 runs …He reached his **half-century** off 68 balls …his maiden **century** in 128 deliveries with the help of six **boundaries** and two sixes.
- **System:** …решающее партнерство **второй калитки** на 93 рана ...своего **полувека** с 68 мячей ...первого **века** за 128 подач с помощью шести **границ** и двух шестерок (cricket terms rendered literally)
- **Reference:** …критически важный результат **у ворот** в 93 партиях ...своего **серединного этапа** за 68 мячей ...первой **сотне** в 128 подач с помощью шести **пограничных** и двух шестерок

A cricket report, a domain with no settled Russian terminology. The system renders *half-century*, *century* and *boundaries* literally; the reference paraphrases them differently and is itself non-idiomatic (серединный этап for *half-century*). Both annotators read a fluent, complete translation and marked one minor span each. The two learned metrics, both reading the hypothesis against this reference, treat the divergence as severe error; BERTScore still credits the shared names, numbers and sentence structure and pulls the composite 12 points closer to the human score. The composite does not repair the failure—both scores are far below 93.5—but it caps the damage a single metric can do, which is the mechanism behind the 69 % figure for "F7 raises" segments.

Case 5—en→ru, literary, system Yandex. Human 100 ($s_1 = s_2 = 100$, no spans). MetricX-XXL = 2.77 (mx_lin = 88.9; below the corpus median of 4.69); xCOMET = 0.52 (bottom fifth of the corpus); BERTScore = 0.80 (corpus median). F7 = 75.3 vs. MetricX-alone 88.9.

- **Source:** He shifted his bag higher on his shoulder, drawn toward **the rink** like muscle memory. …"I'm Declan. **Declan Hull**."
- **System:** Он поправил сумку на плече и, повинуясь мышечной памяти, направился **к катку**. ...— Меня зовут Деклан. **Деклан Халл**.
- **Reference:** Он поправил сумку на плече, инстинктивно направляясь **к хоккейной площадке**. ...— Я Деклан. **Деклан Холл**.

A fluent, faithful literary rendering that both annotators scored 100 without a single span. It diverges from the reference at almost every content word (каток vs. хоккейная площадка, Халл vs. Холл, колледж Святого Кассиана vs. школа), and the reference itself opens a narrative sentence with a stray dialogue dash. MetricX-XXL is nearly right; xCOMET, scoring the same divergence, is in

the bottom fifth of the corpus and pulls the composite down by 13.6 points. This is the coin-flip side of the asymmetry: when the composite lowers a segment relative to MetricX, it is as often penalising legitimate paraphrase as catching an error.

E The exponential F7 variant

The main text uses the linear MetricX-XXL normalisation mx_lin throughout. Here we report the equal-billing exponential alternative

$$\text{mx_exp} = e^{-\text{MetricX-XXL}/3} \cdot 100,$$

together with its rationale and its performance inside F7.

Rationale. The mx_exp form applies a non-linear penalty: segments with few errors (MetricX-XXL $\approx$ 0–2) map to the upper range $[51, 100]$, mid-error segments (5–10) drop sharply to $[4, 19]$, high-error segments ($>$ 12) effectively to zero. This reflects the **non-linear nature of perceived quality**: the gap between "almost flawless" and "minor errors" is qualitatively smaller than the gap between "minor errors" and "many errors". The constant $\tau = 3$ is chosen so that a median "acceptable" translation (MetricX-XXL ≈ 2) gets a score ≈ 51—at the lower edge of the Medium tier. Advantages: (1) *perception-aligned*—the progressive penalty matches the reading experience; (2) *discriminative on the low end*—bad segments map reliably into the Bad zone. Disadvantages: (1) compression at the top of the scale (mx_exp $\in [80, 100]$ corresponds to MetricX-XXL $\in [0, 0.67]$, a narrow band in which excellent and good translations are hard to separate); (2) high MAE when combined linearly with metrics that have a different central tendency.

Empirical comparison of lin vs. exp at the single-metric level. Before plugging the normalised metric into the composite AQI it is worth checking that the normalisation itself does not throw away the signal already present in raw MetricX-XXL against the human score. The anchor is the correlation of the negated raw MetricX-XXL with the human score.

- **Spearman is identical** across the three forms (0.4851). Spearman is computed on ranks, and any monotone function preserves the order of observations.

Form	Pearson	Spearman	Tier 85/60
$-$mx (anchor)	0.4500	0.4851	—
mx_lin	**0.4500**	**0.4851**	**50.0%**
mx_exp($\tau = 3$)	0.4383	0.4851	23.2%

Table 10: Single-metric MetricX-XXL normalisations against the human score, $n = 3{,}450$. Sanity check before plugging into the composite AQI.

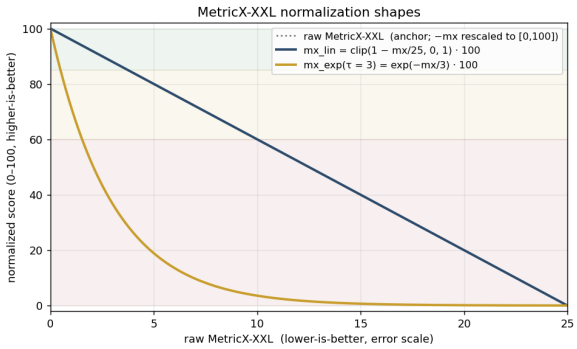

Figure 5: Three MetricX-XXL normalisation curves on $x \in [0, 25]$, $y \in [0, 100]$. Dashed grey—raw MetricX-XXL shown as an affine remap so that it shares the y axis with the two normalised forms (it coincides with mx_lin). Blue—mx_lin. Yellow—mx_exp($\tau = 3$). Bands mark tier zones (red < 60, yellow 60–85, green ≥ 85).

- **Pearson under mx_lin equals the anchor exactly** (0.4500). mx_lin is an affine transform of raw MetricX-XXL, and Pearson is invariant under affine transforms.
- **Pearson under mx_exp($\tau = 3$) is 0.012 lower** (0.4383). The relationship (mx_exp, human) becomes non-linear, and Pearson—a measure of linearity—drops.
- **Tier-agreement** under mx_exp is dramatically worse (23.2% vs. 50.0%): the curve $e^{-\text{mx}/3}$ compresses most segments into the lower part of the scale, and most of them fall into the Bad zone under the 60/85 tier boundaries.

We also tried non-linear normalisations with more free parameters (a τ sweep for exp; a 2-parameter sigmoid; a 4-parameter piecewise-linear), fitting them under 5-fold CV. None gave a meaningful Pearson gain over mx_lin under honest cross-validated evaluation: exp catches up with linear only at large τ, at which the shape is already visually close to a straight line, and sigmoid and piecewise add free parameters without robust

improvement.

F7-exp inside the composite. Substituting mx_exp for mx_lin in F7 (Eq. 2) gives

$$\begin{aligned}\text{AQI}_{\text{exp}} &= 0.40 \cdot e^{-\text{MetricX}/3} \cdot 100 \\ &+ 0.30 \cdot \text{xCOMET} \cdot 100 \\ &+ 0.30 \cdot \text{BERTScore} \cdot 100.\end{aligned}$$

On ranking and linear correlation F7-exp is almost indistinguishable from F7-linear (Spearman $+0.506$ vs. $+0.505$; Pearson $+0.476$ vs. $+0.483$), but its compressed scale pushes most segments into the Bad zone under the 60/85 thresholds, so tier-agreement is far worse (29.4% vs. 46.5%) and MAE much higher (24.33 vs. 14.55) without calibration. Per-LP offsets for F7-exp are all positive (mostly compensating for the compression): cs→de $+20.24$, en→cs $+25.61$, en→es $+28.74$, en→et $+6.63$, en→ru $+19.73$; with them F7-exp reaches Pearson $+0.596$, tier 53.0%, MAE 13.86—converging to the same tier-agreement as F7-linear+LP. Folding in a human score (§4) uses the identical $\alpha = 0.30$ blend; the exp form is identical to the linear hybrid except for the MetricX normalisation.

Ridge coefficients. The unconstrained Ridge regression of §3.4.3 ($\alpha = 1$, standardised features) has coefficients

```
AQI_Ridge = +10.86
          + 0.58 * mx_lin     + 0.24 * xc100
          + 0.11 * bert100    + 0.09 * chrf100
          + 0.07 * ter_inv100 - 0.20 * ck100
```

The negative ck100 weight is the suppressor effect; Ridge-5 drops it and recovers all-positive coefficients within ≈ 0.01 Pearson of the 6-metric fit.

F Geometric-mean composites

Table 11 reports the Geo family discussed in §3.5: Geo3 is the geometric mean of the three normalised F7 candidates; Geo4 and Geo5 add further metrics, Geo5 including TER and chrF.

G Scatter plots for the formula variants

Figure 6 is the 5-fold-CV scatter for the headline F7-linear formula; Figures 7–9 give the same for the non-finalist variants discussed in §3.5.

H Neural text-encoder experiments

We also tried full-blown neural regressors consuming src/ref/tgt text directly—a late-fusion MLP, a ResMLP cross-encoder, and an extended version with error-span features. The last reached a high test $\rho \approx 0.80$, but is disqualified due to **label leakage**: error-span features were extracted from the same ESA span annotations from which the target is built. None of the three fit the "formula on metric outputs, not on text" requirement and they do not compete with F7 as a closed-form composite.

Formula	Pearson	Spear.	Tier	MAE
AQI-Geo3	+0.492	+0.507	28.3%	28.32
AQI-Geo3 + LP	+0.579	+0.583	54.8%	14.65
AQI-Geo4	+0.495	+0.520	27.3%	27.37
AQI-Geo4 + LP	+0.592	+0.601	54.6%	14.22
AQI-Geo5	+0.178	+0.517	24.3%	67.69

Table 11: Reference-based geometric composites. Geo3 and Geo4 are competitive with F7 on Pearson but, without calibration, lose heavily on tier-agreement (28%) due to a compressed distribution. Geo5 with TER and chrF breaks down ($\rho = 0.178$): weak metrics pull the geometric mean down.

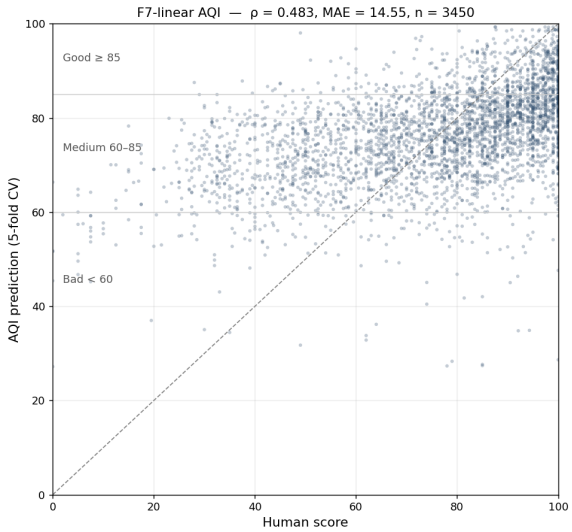

Figure 6: F7-linear AQI vs. human score, 5-fold CV. $\rho = 0.483$, MAE 14.55.

I AQI interpretation scale

The explicit deployment interpretation scale for the uncalibrated F7-linear index referenced in §5.2 is given in Table 13 at the end of the paper.

J Normalisation of the five rescaled metrics

The five metrics other than MetricX-XXL (§2.4) are mapped to $[0, 100]$ higher-is-better by rescaling and, for TER, inversion:

```
xc100      = xCOMET-XXL    * 100
bert100    = BERTScore F1  * 100
chrf100    = chrF
```

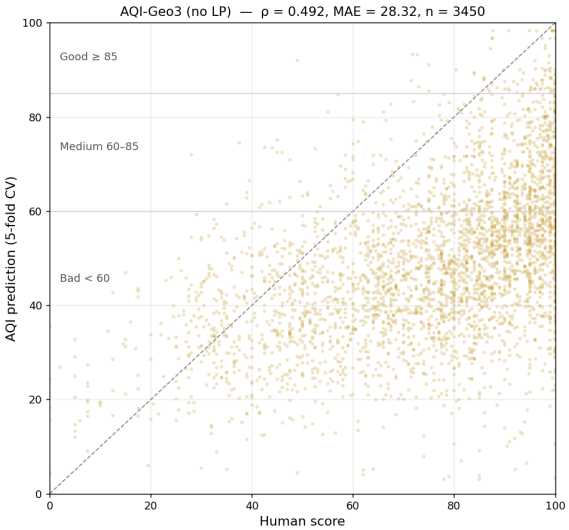

Figure 7: AQI-Geo3 (no LP), 5-fold CV. $\rho = 0.492$, MAE 28.32. Pearson slightly above F7-linear, but the distribution is heavily compressed into the lower band—hence the large MAE and poor tier-agreement without calibration.

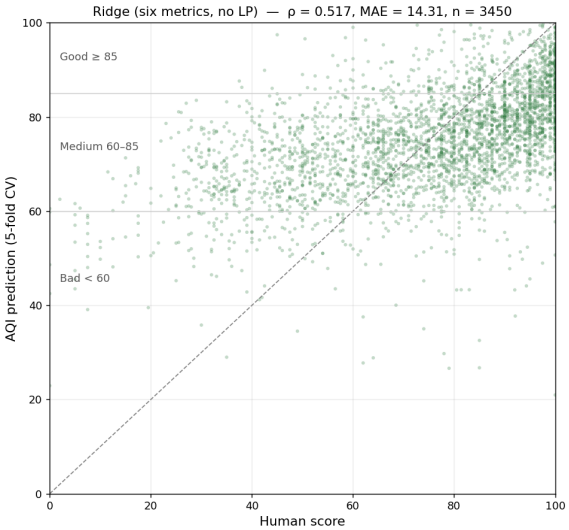

Figure 8: Ridge on six metrics (no LP), 5-fold CV. $\rho = 0.517$, MAE 14.31. Slightly above F7-linear on Pearson, but with a negative COMETKiwi weight (§3.4.3).

Domain	n	F7	MetricX	Refit w	Refit ρ
All	3,450	0.483	0.450	$(0.45, 0.15, 0.40)$	0.490
literary	1,444	0.572	0.579	$(0.60, 0.05, 0.35)$	0.603
news	1,406	0.426	0.458	$(0.90, 0.10, 0.00)$	0.462

Table 12: Pearson ρ with the human score by domain. F7 uses the fixed weights $(0.40, 0.30, 0.30)$; "Refit" $\boldsymbol{w} = (w_{\text{mx}}, w_{\text{xc}}, w_{\text{bert}})$ is the best point of the same 0.05-step simplex grid fitted *in-sample* on that domain, an optimistic upper bound. Cross-domain transfer: literary weights applied to news give $\rho = 0.441$, news weights applied to literary $\rho = 0.583$. Low-quality tail (human < 60): F7 vs. MetricX 0.338 vs. 0.297 on literary, 0.276 vs. 0.282 on news. en→es carries no domain label and is excluded.

```
ter_inv100 = clip(1 - TER/100, 0, 1) * 100
ck100      = COMETKiwi-XXL * 100
```

K Per-domain behaviour

Table 12 gives the per-domain figures behind the "Domain dependence" paragraph of §3.7: F7 with the fixed weights, MetricX-alone, and an in-sample refit of the three weights on each domain, together with cross-domain transfer and low-quality-tail correlations in the caption. The script and CSV are in the repository.

L Future work

(i) **QE-only AQI** for zero-shot reference-free deployment—a separate study dedicated to QE metrics. (ii) **Per-domain calibration** instead of per-LP only, especially for production scenarios with a known domain (legal, medical, marketing). (iii) A **continual learning** protocol: automatic recalibration of the formula on a rotating held-out corpus on each new MetricX-XXL/xCOMET release.

AQI range	Category	Description	Recommended action
≥ 83	Good	**Near-human quality; fluency and adequacy are strong.** This reflects translations that meet near-human quality levels. Little intervention is needed.	Minimal post-editing / review
75–83	Acceptable	**Good fluency, minor errors; meaning is mostly preserved.** This range reflects translations that are good enough for practical use, especially in low-risk scenarios. They likely require some linguistic refinement, but the core meaning is there, making light post-editing efficient.	Light to moderate post-editing
68–75	Medium	**Understandable but contains noticeable issues.** At this level the translation is functional but likely requires a decent amount of intervention to correct grammar, syntax, or terminology.	Moderate post-editing needed to improve fluency
60–68	Low	**Translation is usable with effort; lacks fluency / adequacy.** The output is typically flawed enough that it would need substantial post-editing. Can serve as a rough draft for human editors.	Heavy post-editing required
< 60	Poor	**Unusable or unintelligible translation.** The translation is far enough from acceptable quality that it is not worth post-editing. It is also an indicator that the engine may need re-tuning or re-training for this language pair or content type.	Reject or re-translate; possibly engine tuning

Table 13: AQI interpretation scale for the uncalibrated **F7-linear** index. Breakpoints $60/68/75/83$ are chosen so that the mean human score of the segments in each band ($\approx 88, 79, 70, 61, 53$ from Good to Poor on $n = 3{,}450$) matches the label; band shares are $25.3/33.1/23.1/13.4/5.0\%$. The same labels carry over to the F7+LP variant on its recentred scale (§3.7). For a new LP or domain outside the studied slice, recalibrate the breakpoints.

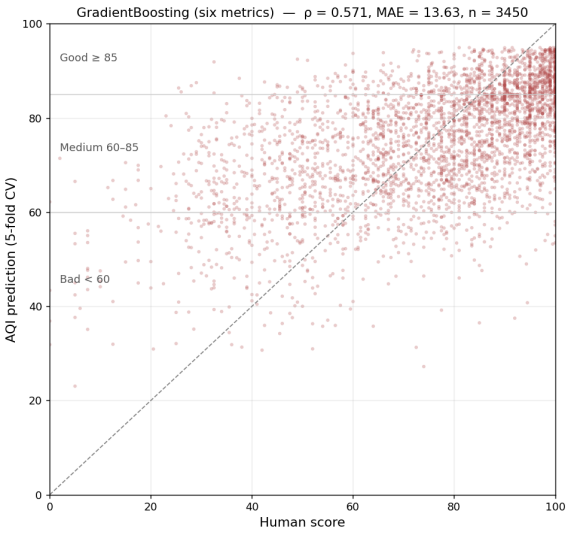

Figure 9: GradientBoosting on six metrics (no LP), 5-fold CV. $\rho = 0.571$, MAE 13.63—the ML ceiling on this feature set. Not a closed-form formula.